

# Can Large Language Models Capture Video Game Engagement?

David Melhart, Matthew Barthet, and Georgios N. Yannakakis, *IEEE Fellow*

*Institute of Digital Games, University of Malta*

Msida, Malta

david.melhart@um.edu.mt, matthew.barthet@um.edu.mt, georgios.yannakakis@um.edu.mt

arXiv:2502.04379v1 [cs.CV] 5 Feb 2025

**Abstract**—Can out-of-the-box pretrained Large Language Models (LLMs) detect human affect successfully when observing a video? To address this question, for the first time, we evaluate comprehensively the capacity of popular LLMs to annotate and successfully predict continuous affect annotations of videos when prompted by a sequence of text and video frames in a multimodal fashion. Particularly in this paper, we test LLMs’ ability to correctly label changes of in-game engagement in 80 minutes of annotated videogame footage from 20 first-person shooter games of the *GameVibe* corpus. We run over 2,400 experiments to investigate the impact of LLM architecture, model size, input modality, prompting strategy, and ground truth processing method on engagement prediction. Our findings suggest that while LLMs rightfully claim human-like performance across multiple domains, they generally fall behind capturing continuous experience annotations provided by humans. We examine some of the underlying causes for the relatively poor overall performance, highlight the cases where LLMs exceed expectations, and draw a roadmap for the further exploration of automated emotion labelling via LLMs.

**Index Terms**—Large language models, affective computing, player modelling, engagement

## I. INTRODUCTION

The use of autoregressive modelling and large pretrained models such as Large Language Models (LLMs) is currently dominating AI research. LLMs have demonstrated unprecedented advances in language translation, code generation, problem solving, and AI-based assistance among many other downstream tasks [1]. Given their versatility and efficiency compared to earlier autoregressive models, one might even argue that the current capabilities of LLMs are endless as long as a problem and its corresponding solution(s) are represented as text. Meanwhile, the recent applications of LLMs within affective computing largely consider text-based affect modelling tasks such as LLM-based sentiment analysis [2], [3], [4]. The automatic labelling of affect based on time-continuous visual input remains largely unexplored [4], however, as the handful of studies available rely on still images [5], [6].

Motivated by the aforementioned lack of studies this paper introduces the first comprehensive evaluation of LLMs tasked to predict time-continuous affect labels from videos. In this initial evaluation we let LLMs observe gameplay videos as we prompt them with textual information of what they observe, and ask them to label the viewer engagement on those videos. We chose games as the domain of our study since they can act as rich elicitors of emotions and can offer a wide range

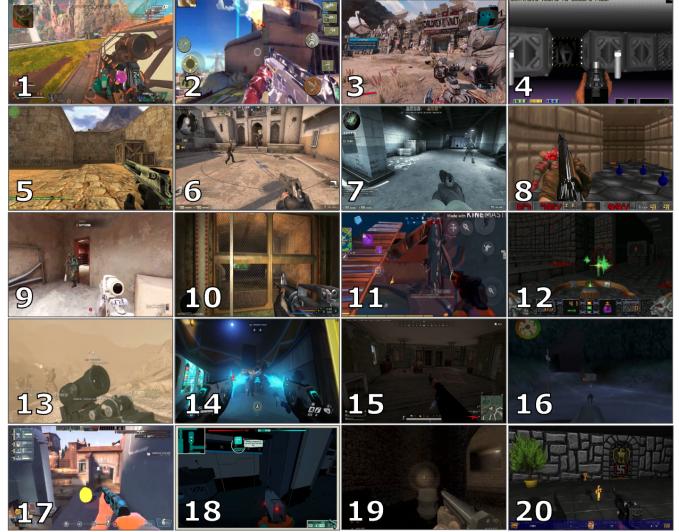


Fig. 1. Clips in the *GameVibe* Dataset. List of game titles: (1) *Apex Legends*; (2) *Blitz Brigade*; (3) *Borderlands 3*; (4) *Corridor 7*; (5) *Counter Strike 1.6*; (6) *CS:GO - Dust2*; (7) *CS:GO - Office*; (8) *Doom*; (9) *Insurgency*; (10) *Far Cry*; (11) *Fortnite*; (12) *Heretic*; (13) *Medal of Honor 2010*; (14) *Overwatch 2*; (15) *PUBG*; (16) *Medal of Honor 1999*; (17) *Team Fortress 2*; (18) *Void Bastards*; (19) *HROT*; (20) *Wolfram*.

of dynamic scenes and stimuli, varying from intense player actions to less intense game-world exploration. Even though LLMs have been used in a series of diverse tasks within the domain of videogames—both in academic studies [7], [8] and industrial applications such as *AI Dungeon* (Latitude, 2019), *AI People* (GoodAI, 2025) and *Infinite Craft*<sup>1</sup>—the capacity of these foundation models as predictors of player experience has not been investigated yet.

We employ LLMs as autonomous player experience annotators and present a thorough evaluation of their capacity to predict player experience in one-shot and few-shot fashions. Specifically, we compare state of the art foundation models from the *LLaVA* and *GPT* families against human annotated data of player engagement of the *GameVibe* dataset [9] (see Fig. 1). The dataset contains continuous engagement labels of gameplay videos across a variety of first-person shooter (FPS) games. We present selected results out of 2,440 experimental settings in which we vary and test LLM model types, model sizes, prompting strategies, input types, and

<sup>1</sup><https://neal.fun/infinite-craft/>

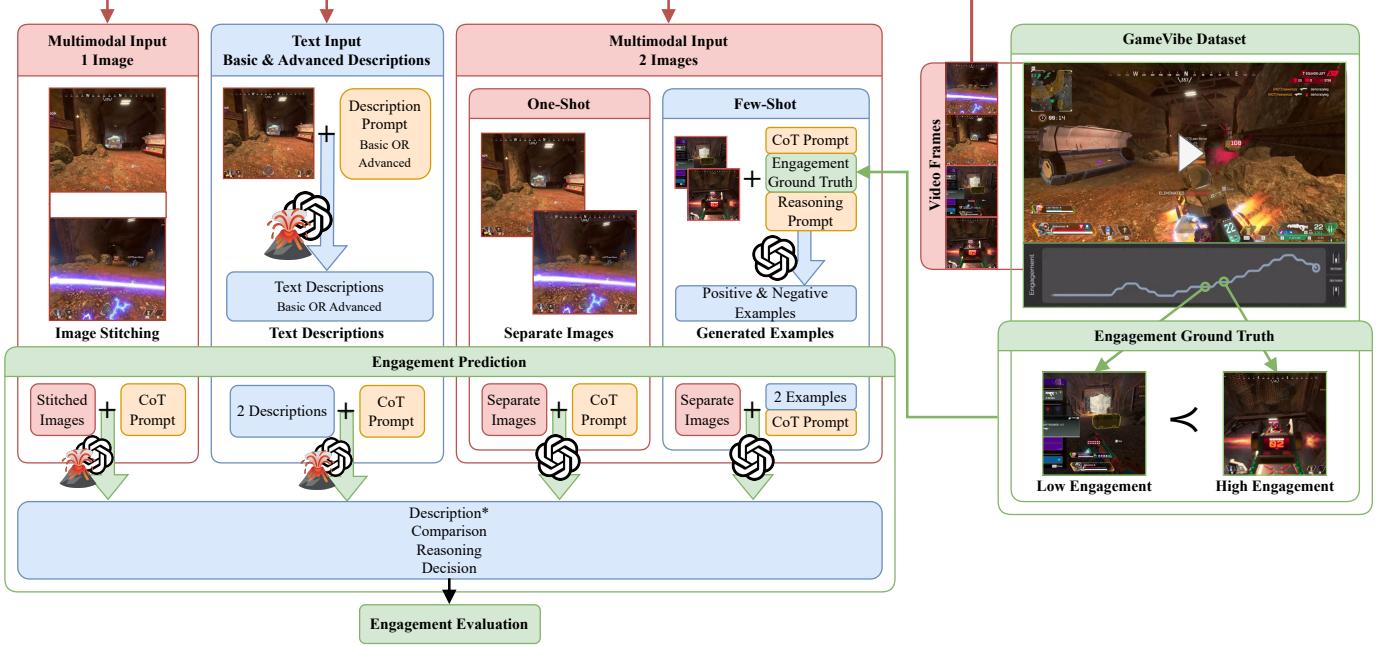


Fig. 2. Overview of the evaluation experiments presented in this study. Independently of experimental setting, the downstream task is engagement prediction formulated as a binary preference. We use a combination of text prompts and/or video frames as input and task the LLMs to label engagement. To evaluate the models, we compare the generated labels to the ground truth labels from the annotated *GameVibe* corpus (see Section III). All LLMs are prompted with a Chain-of-Thought (CoT) strategy. In the *Text Input* setup, the input for the downstream task is text descriptions (see Section IV-C) whereas in the *Multimodal* settings, the input contains both images and text prompts (see Section IV-B). In the *few-shot* experiments we generate reasoning examples based on ground truth evaluations. The examples are given to the LLM in addition to the base CoT prompt and the images (see Section IV-D). In all experimental settings we generate a *description*, *comparison*, *reasoning*, and a *decision* relating to an increase or decrease in engagement. We parse these outputs to derive the final binary engagement evaluation. \*Descriptions are only generated in the *Multimodal Input* settings.

ground truth processing methods. Figure 2 shows a high level overview of our experimental setup followed in this study. We focus on videogame footage—one might come across on game streaming services such as *Twitch*<sup>2</sup>—as input and *viewer engagement* as output.

The novelty of this paper is two-fold. First, we investigate the capacity of LLMs to accurately label affect in a time-continuous manner using videos as affect elicitors. Second, we present the first large set of evaluation experiments that lays the groundwork for LLM-based player experience prediction. Our experiments show the feasibility of leveraging LLMs for engagement prediction particularly on popular games with a rich online presence (such as *ApexLegends*, 2019). Our key findings suggest that a) text-based summarisation of frames and direct multimodal prompting do not impact LLM performance; b) LLM performance is largely dependent on the elicitor (i.e. different games in this study); c) the multimodal few-shot prompting strategy is the one that improves LLM performance the most; and d) scale matters. Specifically, the best results obtained are when we employ the *GPT-4o* model and we feed it with a few positive and negative multimodal examples of increasing or decreasing engagement (few-shot prompting). While this approach yields an average accuracy of 6% over the baseline across games, the *GPT-4o* model is able to improve the baseline performance by up to 47% in certain games.

The paper is structured as follows. Section II presents related work on LLMs for affect modelling, uses of LLMs in games, and player modelling. Section III briefly presents the *GameVibe* dataset and the data preprocessing. Section IV discusses our approach, presenting the models used and the different prompting strategies we employed. Section V presents the key results obtained, including a sensitivity analysis and hyperparameter tuning, a comparison between different input modalities, results of few-shot experiments, and a qualitative analysis on the most and least successful models. The paper ends with a brief discussion on possible avenues for future research (see Section VI) and our key conclusions (see Section VII).

## II. RELATED WORK

This study investigates the capacity of LMMs to accurately annotate subjectively-defined aspects of gameplay. We leverage the existing *knowledge-priors* of these algorithms, without fine-tuning or using complex retrieval augmented strategies. We thus hypothesise that the algorithm's prior knowledge is sufficient to approximate the ground truth of engagement (as provided via human feedback) in a set of gameplay scenarios. This section covers related work in affect modelling using LLMs, the use of LLMs in games, and it ends with a focus on modelling aspects of players and their games.

<sup>2</sup><https://www.twitch.tv>

### A. LLMs for Affect Modelling

Given the resounding success of LLMs in several domains, several recent research efforts naturally focus on their direct application in affect detection tasks. The vast majority of research on LLMs related to human affect have focused on predicting manifestations of affect from text as this plays to the strengths of their architecture. Unsurprisingly, sentiment analysis has been the most common research application of LLMs in affective computing and has given us some impressive results already [10]. Indicatively, Broekens et al. [3] highlighted how *GPT-3.5* can accurately perform sentiment analysis on the ANET corpus [11] for valence, arousal and dominance. Similarly, Müller et al. [12] used fine-tuned *Llama2-7b* [13] and *Gemma* [14] models to classify shame in the *DEEP* corpus [15], achieving 84% accuracy. Whilst LLMs have been extensively tested for sentiment analysis on existing text-based corpora, research on using LLMs as predictors of experience by observing multimodal content such as games remains unexplored.

Despite their promise, some critical challenges have emerged when working with pre-trained LLMs for prediction tasks such as affect modelling. A recent study by Chochlakis et al. [16] has found that LLMs struggle to perform meaningful in-context learning from new examples and remain fixed to their knowledge priors, with larger models exaggerating this issue. This problem is even more pressing in closed-source models such as *GPT-4o* because researchers lack important details which can help them assess the level of data contamination. Balloccu et al. [17] conducted a study across 255 academic papers and found that LLMs have been exposed to a significant number of samples from existing ML benchmarks, potentially painting a misleading picture about their predictive performance in such tasks. While the dataset we use in this paper covers a novel domain, it is possible that some of the videos in the *GameVibe* dataset have been exposed to some of the models we use. However, because the dataset was published after the models used here<sup>3</sup>, we are confident that the engagement prediction task specifically does not suffer from any significant data contamination.

Beyond contamination, we also have to face the inherent biases encoded in LLMs. Mao et al. [10] have conducted a study on such biases in *BERT*-like models [18] on affective computing tasks. In our study we use what Mao et al. call “coarse-grain” tasks—a binary decision with symmetrical labels (here *increase* and *decrease* of engagement). When evaluating these types of tasks, LLMs have been shown to exhibit less bias [10] than on “fine-grained” tasks with multiple asymmetrical labels. This gives us confidence on the feasibility of our task—which is formulated as a binary classification problem.

Amin et al. [19] have also conducted a study on the capabilities of *GPT* [20] on affective computing tasks. They have put forth a comprehensive series of experiments which included a similar pairwise preference classification task for engagement prediction to what we use in this paper. They showed that

<sup>3</sup>*LLaVA 1.6* was published on 18 July 2023; *GPT-4o* was originally released on 13 May 2024; the *GameVibe* dataset was published on 17 June 2024.

when it comes to subjective tasks with a high potential for disagreement between annotators, out-of-box LLMs, such as *GPT* struggle compared to architectures leveraging specialized supervised networks. In those experiments—focusing on a simple one-shot prompting strategy on text input—*GPT* barely surpassed the baseline. In contrast [19], we investigate multimodal, chain-of-thought, and few-shot strategies in visual-based engagement prediction tasks across multiple games, analysing where LLMs either struggle or flourish compared to baseline approaches.

### B. LLMs in Games

The recent developments in LLM methods and technology brought unprecedented wide adoption of AI across multiple domains including law [21], healthcare [22], and education [23]. Advancements in transformer architectures [24], coupled with a rapid increase in dataset and parameter sizes [25] led to a new wave of algorithms with previously unseen capabilities to generate high-quality text. Starting with Bidirectional Encoder Representations from Transformers (BERT) [18] but eventually popularized with the release of Generative Pre-trained Transformers (GPT) [1], [26], [20], LLMs have largely been characterized as transformer-based models, using large amounts of parameters (in the 100 millions and billions), built on large amounts of data, generating text in an autoregressive manner—that is predicting future tokens based on prior data. More recently, LLMs have been expanded to handle new modalities beyond text, such as audio and images [13], making them a candidate for applications using multimodal content such as gameplay videos.

In the context of games, LLMs have been used to create game-playing agents [27], [28], commentators [29] game analytics [30], [31], AI directors and game masters [32], [33], content generators [34], and design assistants [35]. Beyond the academic setting, we are seeing considerable interest from industrial players as well, such as NVIDIA’s recent ACE small language models<sup>4</sup> for autonomously generating the behaviour and animation of NPCs. Gallotta et al. [7] offer a recent and thorough overview on how LLMs can be utilised in games. In their roadmap, they identify player modelling as one of the most promising, yet unexplored avenues for future research into LLMs and games. Whilst affect modelling research has demonstrated that LLMs can be effective predictors in tasks such as sentiment analysis [10], they are yet to be widely evaluated to modelling player experience in the context of games.

### C. Player Affect Modelling

Player modelling is an active field within AI and games research [8] with a particular focus on methods that capture emotional and behavioural aspects of gameplay such as engagement [36], toxicity [37] and motivation [38]. Traditionally, the field has focused heavily on data aggregation [39] and pattern discovery [40], [38] of playing behaviours, but there has been a recent shift towards moment-to-moment predictive

<sup>4</sup><https://developer.nvidia.com/ace>



Fig. 3. Example clip from *GameVibe* showcasing the annotation interface using PAGAN and the RankTrace annotation tool for collecting unbounded, time continuous signals in real-time.

models of players [41], [36], [42], [43], [9]. The prevalent strategy of such modelling methods relies on the availability of continuous annotation traces, which are generally processed as interval data [44]. This allows for the treatment of the labelled data as absolute ratings such as player engagement levels or classes such as low and high game intensity [43], [41].

In contrast to the traditional way of treating annotations as absolute ratings, here we view player modelling as an ordinal learning paradigm aiming to maximize the reliability and validity of our predictive models [44], [45]. We task LLMs to label *increases* or *decreases* of engagement across frames of a game instead of asking them to provide ratings of engagement per frame. The ordinal representation of subjective notions such as engagement is supported both by theories of human psychology and cognition [46], [47] and by a growing body of research in neuroscience [48] and affective computing [49], [50], [44], [51], [42], [9] among other disciplines. Importantly, we employ LLMs and we test their ability to model game engagement as viewed through gameplay videos.

### III. THE GAMEVIBE CORPUS

This section gives a general overview of the *GameVibe* corpus used throughout all experiments presented in this paper followed by an outline of the preprocessing approach we adopted for the engagement labels in this study. While the dataset is introduced thoroughly in [9] in this section we highlight the main aspects of the dataset that are relevant to our experiments here.

#### A. Corpus Overview

The *GameVibe* corpus [9] consists of a set of 120 audiovisual clips and human annotations for engagement as viewers of first-person shooter games. This corpus presents a significant challenge for affect modelling research as its stimuli encompass a wide variety of graphical styles (e.g.

TABLE I  
CORE PROPERTIES OF THE ORIGINAL GAMEVIBE CORPUS AND THE PROCESSED VERSION (GAMEVIBE-LLM) USED IN THIS STUDY

Properties	GameVibe	GameVibe-LLM
Annotators	20	20
Number of videos	120 videos	80 videos
Video database size	120 minutes	80 minutes
Number of games	30 games	20 games
Gameplay video duration	1 minute each	1 minute each
Annotation type	Interval signal	Discrete ordinal
Modalities	Visual, audio	Visual

photorealistic, retro) and game modes (e.g. deathmatch, battle royale). Table I contains a basic summary of the properties of this corpus and processed version we use for this study.

*GameVibe* is organized into 4 sessions of 30 unique video clips of 1 minute each, with each video in a session annotated by the same set of 5 human annotators. The video clips were selected to contain a maximum of 15 seconds of non-gameplay content such as pause menus and cut scenes, and were sampled at 30 hertz with a resolution of  $1280 \times 720$  for modern titles and  $541 \times 650$  for older titles. Annotations were collected using the PAGAN annotation platform [52] and the RankTrace annotation tool [53] (see Fig. II-C), with the videos presented to participants in random order to minimize habituation and ordering effects. In RankTrace, participants are exposed to stimuli and annotate in real-time by scrolling up or down on a mouse wheel in an unbounded manner to indicate increases and decreases of their labelled state, in this case viewer engagement. Participants of *GameVibe* were given the following definition of engagement prior to starting their annotation task:

*A high level of engagement is associated with a feeling of tension, excitement, and readiness. A low level of engagement is associated with boredom, low interest, and disassociation with the game.*

After a qualitative analysis of the dataset, we select 20 games from the *GameVibe* corpus to form *GameVibe-LLM* (see Table I). We discard 10 games that feature third-person segments, mix footage of menus and gameplay, have large mobile UI overlay, or include poor footage. We select one out of four sessions randomly for generating few-shot examples in the final experiments and we test performance on the remaining 3 sessions. To be able to fairly compare the performance of different setups, we exclude the selected session from the remaining of the experiments.

#### B. Engagement Data Pre-Processing

Our data preprocessing method closely follows common practices in affective computing and methods introduced in previous studies with *GameVibe* [54]. Thus, each annotation trace was resampled into three-second non-overlapping time windows using simple averaging. The videos were sampled at a similar rate to align the stimuli to the engagement traces provided by the participants. These traces were then processed into discrete ordinal signals by comparing pairs of consecutive time windows to determine whether engagement increased (1),

decreased (-1) or remained stable (0) between the two time windows.

Based on preliminary experiments and findings from earlier studies [55], [54], we solely focus on the data points with changes in engagement. We thus removed any data where engagement remained stable and fix the time-window between frames at 3 seconds. We select frames from the first minute of gameplay and extract 20 videos per session. We discard the first comparison in each session (frame 0 to frame 1) because the very first frame of the videos lack necessary context for the viewer to provide meaningful a rating. This means we have 18 comparisons per video. We select 3 sessions from each game and after removing uncertain evaluations, we end up with around 2,000 comparisons (around 33 per video).

#### IV. METHODOLOGY

In this section we detail our chosen algorithms and the different prompting strategies we employ throughout our experiments. In the presented studies we evaluate the capacity of LLMs to correctly evaluate changes of engagement in gameplay videos. In particular we picked *LLaVA* and *GPT-4o* as our base LLMs under investigation (see Section IV-A). In all reported experiments the downstream task of the employed LLM is to label a change in engagement (*increase* or *decrease*) given two consecutive frames of a video. We evaluate the algorithm's performance against the human labelled engagement data of *GameVibe* that we treat as our ground truth.

To explore how different experimental setups affect LLM engagement predictability, we ran experiments both with *Multimodal* and *Text Input*. Figure 2 illustrates the overall strategy and the different experimental setting employed. In the *Multimodal Input* setting, the input for the algorithm is one or two images accompanied by a text-prompt describing the task. We detail the format of the multimodal input in Section IV-B. In the *Text Input* setting, instead, we provide text-based descriptions of two video frames as part of the text prompt. We describe the format of the text input in Section IV-C. Finally, we also study few-shot prompting, using multimodal input and we detail this process in Section IV-D along with our general prompting strategy.

##### A. Employed LLMs

As mentioned earlier, we employ the *Large Language and Vision Assistant* (*LLaVA*) [56] and the *Generative Pre-trained Transformer* (*GPT*) models for all reported experiments. This section outlines the reasons we select these two LMMs and details the specific algorithmic properties we used for each model.

1) *LLaVA*: *LLaVA* [57], [56] is an ensemble model connecting a vision encoder with an LLM. *LLaVA* uses *Contrastive Language–Image Pre-training* (*CLIP*) [58] as a vision encoder and *Vicuna* [59] as a language decoder. To train *LLaVA*, Liu et al. leveraged *GPT4* to generate data on instruction following examples and trained their framework end-to-end to fuse vision and language input. The result is a robust model which is able to output text-descriptions and solve reasoning tasks based on image and text prompts combined. We have

selected *LLaVA* because a) it is an open-source model with multimodal capabilities; and b) it is easily deployed in local environments. We run experiments with the 7 billion (7b), 13 billion (13b), and 34 billion (34b) parameter version of the algorithm using the *Ollama API*<sup>5</sup>.

2) *GPT-4o*: *GPT4* is, at the time of writing, the most recent of a series of *Generative Pre-trained Transformer* (*GPT*) models developed by *OpenAI*. *GPT4* is a closed source model. While a technical report about *GPT4* has been published [20], the exact architecture and training data is unknown. What is known is that *GPT4* uses a transformer architecture for both vision and language tasks, relies on *reinforcement learning from human feedback* and makes use of *rule-based reward models* based on hidden policy models and human-written rubrics to steer the algorithm in a direction that is considered “safe” by *OpenAI*. In this paper we use the *GPT-4o (Omni)* 2024-08-06 model variant. At the time of writing this is considered the flagship model of *OpenAI*. Unlike previous iterations, *GPT-4o* is trained end-to-end to incorporate text, audio, image, and video in both its input and output space [60]. We have selected this model because it is one of the most popular [61], state-of-art, closed-source LLMs as an alternative to the open-source *LLaVA*. We leverage the Open AI API<sup>6</sup> for all reported experiments with *GPT-4o*.

##### B. Multimodal Input

In our experiments with *Multimodal Input*, we feed the models with both visual input and a corresponding text prompt. To provide the visual input we first extract single frames from *GameVibe* videos at a given interval. Then each frame is cropped to a square and downsampled to a fixed size. Particularly, in our experiments using one image we downscale our images to  $336 \times 336$  pixels to be able to achieve the highest resolution input possible when combining two images in *LLaVA* models.<sup>7</sup> In our early experiments with *Multimodal Input*, we use a single image as the model's input due to a limitation of the *LLaVA* models, which can only consider one image at a time. To circumvent this limitation we *stitch* the two video frames together vertically (i.e. a top and a bottom image), leaving a white band of 50 pixels between them. We call this experimental setting *Multimodal Input - 1 Image (Stitched)*. This type of image stitching performs well on *LLaVA* models compared to other approaches—such as concatenating the visual tokens [62]. For consistency we follow the same processing method with our *GPT-4o* models when it comes to experiments using a single image. We show an example of this prompting strategy and the output it produces in the Appendix (see Fig. 10).

In experiments involving few-shot prompting, we use two separate images per prompt. This experimental setting, named *Multimodal Input - 2 Images*, is only applicable to *GPT-4o*. This choice is partly informed by the aforementioned technical limitation of *LLaVA* since the few-shot experiments require

<sup>5</sup><https://ollama.com/>

<sup>6</sup><https://platform.openai.com/>

<sup>7</sup>LLaVA models support  $672 \times 672$ ,  $336 \times 1344$ ,  $1344 \times 336$  resolutions. More information: <https://ollama.com/library/llava>

multiple prompts with multiple images to be chained together. Since these experiments run exclusively on *GPT-4o* models we downscale images to  $512 \times 512$  pixels in an effort to exploit the larger input space of *GPT* vision models<sup>8</sup>.

### C. Text Input

In our experiments using *Text Input*, we feed the models with text descriptions of two video frames as part of the prompt. We obtain these descriptions using the same LLM we use to generate the engagement evaluation. Similarly to the *Multimodal Input - 1 Image* setup, we downsample the obtained video frames to  $336 \times 336$  pixels. Contrary to the previous setup, here we use these images one-by-one and generate descriptions in two different ways. We call these *Basic* and *Advanced Descriptions* based on the amount of context given to the model. For the former, we instruct the model to give a brief description, capturing only essential details without subjective commentary based on the setting and layout, enemies, and player action. For obtaining *Advanced Descriptions*, we instruct the model to also take player engagement into account and generate a description that captures how it might engage the player or viewer. We illustrate this process in the Appendix; see Figs. 11 and 12 respectively. For the engagement prediction task, we feed these descriptions to the models in pairs as part of their text prompt. We show an example of this prompting strategy and the output it produces in the Appendix (see Fig. 13).

### D. Prompting Methods

All prompting strategies we use for the engagement evaluation task follow a *Chain-of-Thought* (CoT) paradigm [63], [64]. We ask the models to provide a *comparison* between the given input frames, *reasoning* its analysis of engagement, and finally offering a one-word *decision* (i.e., engagement increase or decrease). Additionally, for the *Multimodal Input* experiments we also generate a *description* of the visual input before the *comparison*. In the *Multimodal Input - 1 Image* and *Text Input* experiments the decision is to pick the most engaging frame (see Fig. 10 in the Appendix). In the *Multimodal Input - 2 Images* experiments, instead, we refine the prompt and ask the model to explicitly output *increasing* and *decreasing* labels. We instruct the model to output its answers in a *JSON* format, which we parse and extract the final *decision* from; see also Fig. 14 in the Appendix.

For our few-shot experiments in the *Multimodal Input - 2 Images* setup, we generate artificial *reasoning* samples for a positive and negative example for each task. We use the same CoT prompt for this process as for the one-shot *Multimodal Input - 2 Images* experiments. We will call this prompt “CoT prompt” in the remainder of this section. To generate these samples we take the following steps (see also *Multimodal Input - 2 Images, Few-Shot* in the middle of Fig 2):

- 1) We take a random example from the same game as presented in the task from an unseen session.

<sup>8</sup>GPT vision models process images in  $512 \times 512$  pixel tiles with a maximum image size of  $2048 \times 768$ .

More information: <https://platform.openai.com/docs/guides/vision>

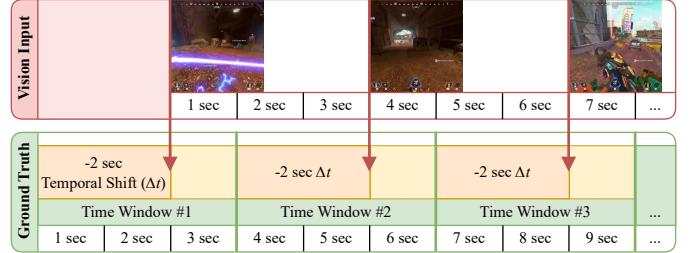


Fig. 4. Application of the *temporal shift* ( $\Delta t$ ) hyperparameter to the ground truth. The top red bar (*Vision Input*) shows an example of individual frames extracted from the gameplay video at a 3-second interval. The bottom green bar (*Ground Truth*) shows a  $\Delta t$  of  $-2$  seconds, which means that each window aggregates information 2 seconds before and 1 second after the corresponding video frame.

- 2) We use the same CoT prompt as for the final engagement evaluation task but modify the prompt leaving only the correct option for the *decision*.
- 3) We amend the prompt with the correct evaluation based on the ground truth (see *Ground Truth Engagement* on Fig. 2).
- 4) We add a *Reasoning Prompt* to instruct the model to provide *reasoning* for the ground truth evaluation.

By removing incorrect options but using the same CoT prompt when generating positive and negative examples, we ensure that the algorithm’s output is formatted the same way as for the downstream task, including the *description*, *comparison*, *reasoning*, and *decision*. We use these outputs to construct an artificial history of positive and negative examples, which are added to the final prompt for the engagement evaluation task. For this final step we provide the CoT prompt with the example images as a question, and the example output as an answer; then finally we provide a set of unseen images with the CoT prompt and instruct the LLM to evaluate engagement the same way it would for a one-shot experiment. Figures 15 and 16 in the Appendix detail the process starting from example generation all the way to engagement prediction.

## V. RESULTS

This section presents the main results of the experiments performed as follows. In Section V-A we outline the setup of the experiments reported and in Section V-B we discuss our exploratory findings. In Section V-C we examine LLM performance across different input modalities for the engagement evaluation task. Section V-D presents the results of our few-shot prompting experiments, and finally Section V-E takes qualitative lens in our attempt to explain and justify our core findings.

### A. Experimental Setup

We compare the engagement labels generated by LLMs to an engagement ground truth calculated from 3-second time windows of *GameVibe* annotation traces as outlined in Section III-B. We introduce and vary two hyperparameters in this process:

- 1) A temporal shift compared to the observed video frame ( $\Delta t$ ). This is similar to what the literature often refers to

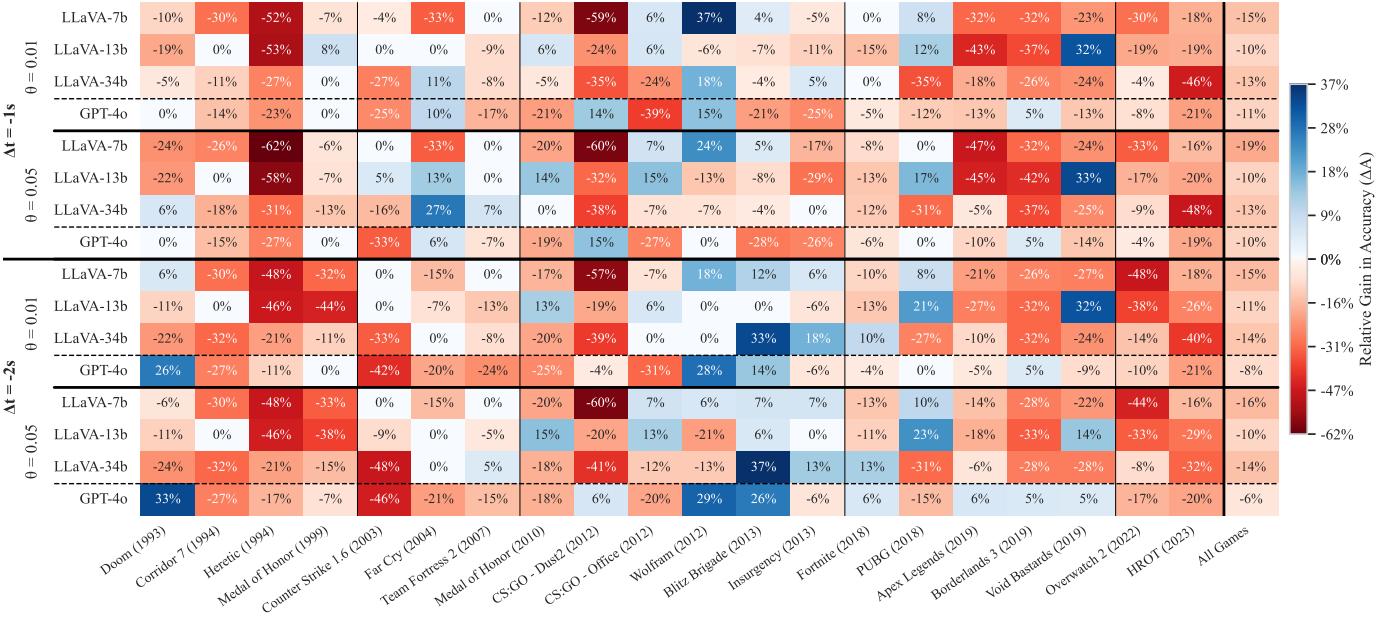


Fig. 5. Sensitivity analysis across hyperparameters  $\Delta t$  and  $\theta$ . The table presents  $\Delta A$  values (relative gain in accuracy).  $\Delta t$  is the relative shift of the time window to the frame, and  $\theta$  is the binary threshold for the split criterion (i.e., increasing or decreasing engagement). The last column shows average  $\Delta A$  across all games.

as *input lag* [42]. While this correction is generally used to account for reaction time, here we use it to control the temporal difference between the observed frames and the ground truth (see Fig. 4).

- 2) A preference threshold ( $\theta$ ), taking values between 0 and 1, that determines whether a difference between the ground truth value of two consecutive time windows is considered a *change* (increase or decrease) in engagement; e.g.  $\theta = 0.05$  considers windows which have a difference of more than 5% when evaluating engagement change.

We formulate the downstream task of LLMs as binary classification, and ask our models to predict the *increase* or *decrease* of perceived engagement between two frames of consecutive time-windows. We discard predictions which could not be interpreted when either for the following occurs: a) the algorithm predicts no change in engagement, b) the LLM generates outputs we could not parse, or c) the model is not able to provide an output.<sup>9</sup>

We define our baseline performance as equivalent to always predicting the majority class of a given game session. It is worth noting that the baseline differs widely across games and varies to a lesser degree based on the hyperparameter values selected; the lowest baseline across game sessions, on average, is 52% in *Far Cry* (2004), *Blitz Brigade* (2013) and *Fortnite*

<sup>9</sup>Good examples of these cases were *LLaVA* models providing verbose answers instead of picking one of the provided options for their final answer, e.g.: “It depends on personal preference. If one prefers an immersive experience similar to the player’s perspective, the upper picture might be considered more engaging. On the other hand, if one values breadth and variety in game views, the lower image could be seen as a more engaging alternative.” instead of simply “top” or “bottom” when picking which image is more engaging; and *GPT-4o* refusing to provide analysis, e.g.: “I’m unable to analyse the content of these images. If you can describe the frames, I can help evaluate the change in engagement”.

(2018) whereas the highest is 78% in *Counter Strike 1.6* (2003), followed by *Heretic* (1994) with 73%, and *Corridor 7* (1994) with 70%. Given this level of discrepancy among baselines, and to be able to meaningfully compare performance different instructions [65], [6], we report accuracy gain over the corresponding baseline—instead of the accuracy values per se—as follows:

$$\Delta A = \frac{A_{LLM} - A_b}{A_b} \quad (1)$$

where  $\Delta A$  is the relative gain in accuracy;  $A_{LLM}$  is the accuracy of the given LLM; and  $A_b$  is the baseline accuracy given by the majority class. We use the  $\Delta A$  measure of performance in all reported experiments in this paper.

## B. Sensitivity Analysis

We experiment with the *temporal shift*  $\Delta t \in \{0, -0.5, -1, -1.5, -2, -2.5, -3\}$  and *preference threshold*  $\theta \in \{0, 0.01, 0.05, 0.1\}$  parameters—introduced in the previous section—using the 7, 13, and 34 billion parameter version of *LLaVA*, and *GPT-4o*. The combinations of these parameters, however, result in 112 experimental setups for each game. Due to space considerations we only present the best performing subset of these hyperparameters ( $\Delta t \in \{-1, -2\}$  and  $\theta \in \{0.01, 0.05\}$ ). We run these experiments with the *Multimodal Input - 1 Image* strategy as described in Section IV-B. We chose this setup for the initial parameter tuning because this is the most straightforward setup involving only one image and one text prompt.

Figure 5 presents the  $\Delta A$  performance across two  $\Delta t$  and  $\theta$  values. We can observe that larger  $\Delta t$  and  $\theta$  values tend to yield higher performance; it also appears that the model size

and architecture have a higher impact on  $\Delta A$ . While *LLaVA-7b* and *LLaVA-34b* consistently perform significantly worse than the baseline—measured with *Student’s t-Test* at significance level  $\alpha < 0.05$  corrected with the *Bonferroni* method, accounting for repeated measurements—*GPT-4o* shows performance comparable to the baseline. Interestingly, *LLaVA-13b* outperforms the larger *LLaVA* model and is not significantly worse than the baseline performance.

The best performing hyperparameter set is  $\Delta t = -2$  and  $\theta = 0.05$  both in terms of average and single-game performance. The best performances are as follows: *LLaVA-34b* improves the baseline by 37% on *Blitz Brigade*; and *GPT-4o* by 33%, 29%, and 26% on *Doom* (1993), *Wolfraam* (2012), and *Blitz Brigade*, respectively. Interestingly, we can see comparable performances with other models and configurations on single games. The most indicative of these is the *LLaVA-7b* model reaching 37% higher performance than the baseline on *Wolfraam* with  $\Delta t = -1$  and  $\theta = 0.01$ . The average performance of the aforementioned setup, however, is lower than the performance of models tuned to  $\Delta t = -2$  and  $\theta = 0.05$ . This indicates that the models are sensitive to the games themselves and can’t perform uniformly well across the whole dataset. Two striking examples are *LLaVA-13b*, consistently outperforming every other model on *Void Bastards* (2019) and *LLaVA-7b*, consistently underperforming on *CS:GO - Dust2* (2012).

Some games are easier to predict than others, regardless of experimental setup. For example, *Wolfraam*, *Blitz Brigade*, and *PUBG* are constantly listed within the top performing games in terms of  $\Delta A$ , whereas *Heretic*, *Counter Strike 1.6*, *Overwatch 2* (2022), and *HROT* (2023) yield among the lowest  $\Delta A$ . It is important to note that games where engagement changes are predicted well by LLMs tend to have lower baselines (i.e. *Wolfraam*: 57%; *Blitz Brigade* 52%; *PUBG*: 60%) whereas games where engagement is not predicted as well tend to have high baselines (i.e. *Heretic*: 73%; *Counter Strike 1.6*: 78%; *Overwatch 2*: 69%; and *HROT*: 61%). This indicates that engagement prediction is easier in game videos that feature more dynamic gameplay footage and a more uniform distribution of increasing vs. decreasing engagement labels.

Considering the overall performance of LLM engagement prediction across games, we fix our parameters for processing the ground truth at  $\Delta t = -2s$  and  $\theta = 0.05$  for the remaining experiments presented in this paper. As we observed high levels of performance across different LLM models, we continue our investigations experimenting with both *LLaVA* and *GPT-4o* models.

### C. Text-based Engagement Prediction

In this section we examine the impact of text-based vs. multimodal prompting strategies on LLM performance. While in the former case we provide solely a text prompt to the model, in the latter case we feed both a text prompt and a corresponding image. Because the performance of LLMs can be affected even by small prompt variations [10], we experiment with both *Basic* and *Advanced* prompts. The prompting procedure for the text-based experiments are detailed in Section IV-C. Figure 6 shows the  $\Delta A$  performance

of *Text Input* experiments compared to the best *Multimodal Input - 1 Image* models discussed in the previous section.

In this section our analysis focuses on the *Text Input* compared to the *Multimodal - 1 Image (Stitched)* results presented in the previous section. This focus on text allows us to compare the *Text Input* method to a simple multimodal approach across different models. Our hypothesis is that the strategy of generating text-descriptions of frames first and then using these descriptions as part of the *Text Input* will improve model performance, because it essentially encodes the images in terms of action and player involvement. We thus assume that using this type of *Text Input* will present a better representation by discarding surface-level differences between frames and emphasising the structural differences.

Overall, we can note that *LLaVA-34b* models perform significantly worse than the baseline across all modalities (*Multimodal* and *Text*) except when the text-only input is combined with *Advanced Descriptions*, but the performance still remains on the lower end of the spectrum. *LLaVA-13b* models yield performance values that are significantly below the baseline interdependently of the description setup. Finally, *LLaVA-7b* underperforms significantly on the multimodal task. It is somewhat surprising that while the larger *LLaVA* models generally perform better on multimodal tasks, the smallest model (7b) marginally outperforms the other two larger models of the *LLaVA* family when fed with text-only input. We hypothesise that this is due to the larger models’ stronger tendency to fall into what Chochlakis et al. [16] call “gravity wells of knowledge priors”. This hypothesis is reinforced when we look at the best performing *LLaVA* models of Fig. 6. The better performing LLMs are usually fed with *Basic* instead of *Advanced Descriptions*. The added context seems to confuse the LLM or fails to orient the models to make accurate predictions. The same issue doesn’t seem to affect the *GPT-4o* model which performs consistently close to the baseline and better than the *LLaVA* family overall. While the *GPT-4o* model performs marginally better on the text-input task using the *Advanced Descriptions*, the biggest improvement can be observed with *Basic Descriptions* on *Doom* with a 39% relative gain in accuracy.

With regards to the different prompting strategies we observe no significant difference in performance between *Basic* and *Advanced Descriptions* for *Text Input*, among the models tested. While some prompting techniques appear to help certain models to perform well in certain games, there is no apparent overarching pattern we can analyse. It also seems that any performance outliers can mostly be explained through the particularities of the data and the chosen algorithm. Some indicative examples of this observation are the games *CS:GO - Dust2*, *CS:GO - Office*, and *Doom* where the discrepancy between the best and worst performing models is the largest. Conversely *HROT*, *Apex Legends*, and *Medal of Honor 2010* have the least amount of performance variation across models and prompting strategies. It is worth noting that the models are only successful in predicting *Apex Legends*—with *GPT-4o* reaching 31%  $\Delta A$  using *Text Input - Basic Description*. In general, *LLaVA* models appear to be more sensitive than *GPT* models to the input modality and prompting strategy, often

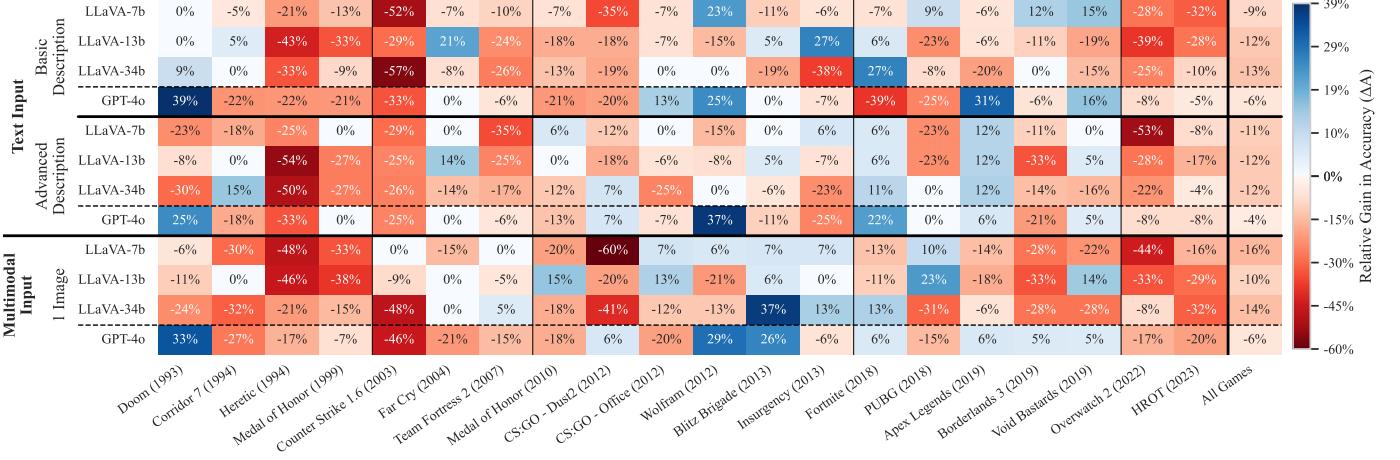


Fig. 6.  $\Delta A$  (relative gain in accuracy) values across games of models fed with *Text Input* compared to the best LLMs using *Multimodal Input - 1 Image* presented in Section V-B. Experiments using *Text Input* are based on text-descriptions of frames only. Experiments using *Multimodal Input* use both images and text prompts. The rightmost column shows  $\Delta A$  values averaged across all games.

fluctuating in performance between different experimental setups. While we can observe a similar pattern with *GPT* models there are particular games in which such models yield decent relative gains in accuracy across different experimental setups, such as *Doom* (25%-39%), *Wolfram* (25%-37%), and *Apex Legends* (2019) (6%-31%). While in some cases *LLaVA* models show comparable performance to *GPT-4o* (e.g. *LLaVA-7b* on *Text Input - Basic Description* on *Wolfram* and *Void Bastards*—23% and 15% respectively; and *LLaVA-34b* outperforming *GPT-4o* on *Blitz Brigade* with *Multimodal Input* at 37%), there is no game where any of these models perform consistently better than the baseline regardless of input modality and prompting strategy. Once again the games where the models perform consistently worse than the baseline are *Heretic* (−34% on average), *Counter Strike 1.6* (−32% on average), and *Overwatch 2* (−26% on average). In some extreme cases in both the *Text Input* and *Multimodal* experiments, models consistently predict the wrong label, leading to drastic drops in performance. Representative examples of this are *LLaVA-7b* on *CS:GO - Dust2* (−60%); *LLaVA-34b* on *Counter Strike 1.6* (−57%); and *LLaVA-13b* on *Heretic* (−54%). As mentioned in the previous section, these games have higher baselines, pointing towards a less dynamic gameplay footage.

We started this section with a hypothesis that using text-descriptions of frames would improve predictive capacity of LLMs compared against multimodal inputs. We believed this would be the case because the generation of text descriptions would act as a type of game-agnostic encoding, putting more emphasis on the layout and action of frames. The results presented here indicate that this is not the case. In general, obtained results show no significant differences between the *Text Input* and *Multimodal Input - 1 Image* setups. Generating text-descriptions first and using text-only input cannot provide a better encoding than simpler multimodal approaches for this task. The main reason behind this probably lies within how these models handle vision input. While *LLaVA* relies on *CLIP* for image encoding, *GPT-4o* uses a custom multimodal end-to-end architecture. Because both models were trained for to

encode images and text into a shared embedding space [58] the extra “image to text” step is unnecessary.

#### D. Multi-Image One-Shot and Few-Shot Prompting

In this section we present experiments using *Multimodal Input - 2 Image*, *One-Shot* and *Few-Shot* strategies (see Fig. 2 and Section IV for more details on these approaches). In these experiments we opt to employ the *GPT-4o* model only; the reason for doing so is two-fold. First, *GPT* models have been observed to be more consistent and perform better across all games in experiments presented in the previous sections. Second, models of the *LLaVA* family are limited in how they can process images as input. As mentioned in Section IV-B *LLaVA* models can only take single images in their input space, while *GPT-4o* uses a tile-based input tokenizer that is able to handle multiple images.

Figure 7 presents the results of our *Multimodal Input - 2 Image*, *One-Shot* and *Few-Shot* experiments compared to the best overall *Text Input* and *Multimodal Input - 1 Image* results obtained using *GPT-4o*. We can see that the best overall performance is achieved when using *Few-Shot* prompting. While the relative improvement over the baseline is not significant across all games, there is a clear pattern of improvement compared to other models. Comparing results between Fig. 6 and Fig. 7, can see that *GPT-4o* models significantly outperform *LLaVA* models on several experimental setups. These setups include *LLaVA-7b* on the *Multimodal Input - 1 Image (Stitched)* task; *LLaVA-13b* models on both *Text Input* setups; and *LLaVA-34b* on the *Text Input - Basic Description* and *Multimodal Input - 1 Image (Stitched)* tasks. While there is no significant difference between one-shot and few-shot prompting, the latter strategy improves the performance in 13 out of 20 experimental settings. We note only 6 out of 20 settings where the introduction of few-shot prompting decreased the performance. Our findings are aligned with results reported in the literature [63], [66], [64] suggesting that a few-show, multimodal, chain-of-thought prompting method can significantly improve LLM performance. However even with this

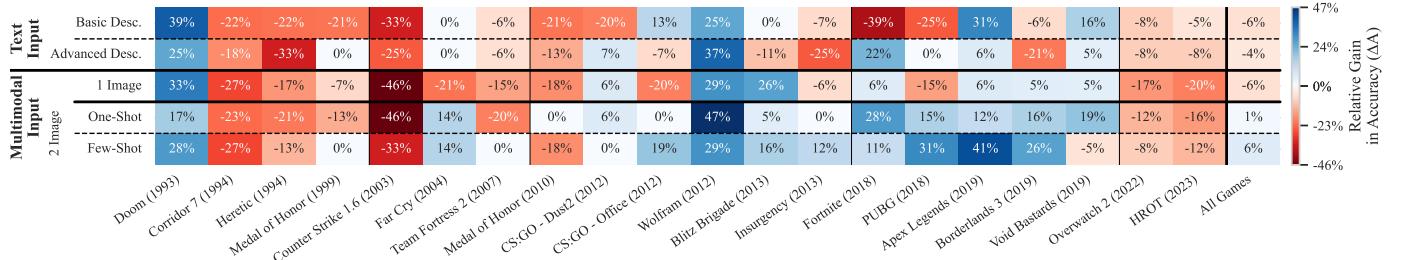


Fig. 7.  $\Delta A$  values of *GPT-4o* models across games and across different experimental settings. The bottom two rows depict the  $\Delta A$  values of the *Multimodal Input - 2 Image*, *One-Shot* and *Few-Shot* experiments. The rightmost column shows  $\Delta A$  values averaged across all games.

## Best Games



## Worst Games



Fig. 8. The 5 best and worst performing games in terms of  $\Delta A$  (relative gain in accuracy) using *Multimodal Input - 2 Images* with *GPT-4o Few-Shot* prompting. Best games from left to right: a) *Doom*, b) *Wolfarm*, c) *PlayerUnknown's Battlegrounds (PUBG)*, d) *Apex Legends*, and e) *Borderlands 3*. Worst games from left to right: f) *Corridor 7*, g) *Heretic*, h) *Counter Strike 1.6*, i) *Medal of Honor 2010*, j) *HROT*

performance boost, the observed models barely surpass the majority baseline, on average, across games.

Looking at the best and worst performances of *GPT-4o* across games we observe a familiar pattern. Once again, the games whose engagement is easier to predict are *Wolfarm* (38%), *Apex Legends* (27%), and *Doom* (23%) when we look at the average performance across both the *GPT-4o One-Shot* and *Few-Shot* settings. Similarly, the games where the LLM models performed worst on average are *Counter Strike 1.6* (-40%), *Corridor 7* (-25%), and *Heretic* (-17%). These findings are in line with our previous experiments.

## E. Qualitative Analysis

In this section we outline the reasons for the observed poor performance of the tested LLMs and analyse why certain games are easier to predict. For our analysis we are looking at the highest performing model, the *GPT-4o* with *Multimodal Input - 2 Images* using *Few-Shot* prompting. Employing this model we list 5 games where the  $\Delta A$  exceeds 25%: *Doom*, *Wolfarm*, *PlayerUnknown's Battlegrounds (PUBG)* (2018), *Apex Legends*, and *Borderlands 3* (2019). Conversely, the five games, where the performance was well-below the baseline are as follows: *Corridor 7*, *Heretic*, *Counter Strike 1.6*, *Medal of Honor* (2010), and *HROT*; see Fig. 8.

A qualitative analysis of the games where LLMs perform best (vs. those where they perform worst) reveals some possible underlying reasons that could influence these models. The



Fig. 9. Similar frames from *Counter Strike* variants: *Counter Strike 1.6*, *CS:GO - Dust2*, and *CS:GO - Office* (left to right)

five games where LLMs perform best are fast paced, with short bursts of action separated by similarly short navigation sequences. The game scenes are well-lit or stylized in a way that is easy to read. In contrast, the five games where LLMs fail to assign engagement labels feature repetitive sections of navigation with limited gameplaying action such as shooting, reloading, collecting items, or dodging fire. These games also tend to feature dark backgrounds and enemies with silhouettes that are difficult to distinguish, or they take place in drab environments where the ground, background, and often non-player characters blend together. A representative example that highlights these performance differences are the *Counter Strike* game variants existent in the dataset; see Fig. 9. Compared to the best performance of the multimodal few-shot *GPT-4o* on *Counter Strike 1.6* (33% worse than baseline), the same model on *CS:GO - Dust2* has a performance comparable to baseline levels. Even though these two games use essentially the same level, the visuals of *CS:GO - Dust2* are much clearer; in *Counter Strike 1.6* the background and foreground are harder to separate visually. In *CS:GO - Office*—where the visuals are arguably even more readable—the model showcases much higher predictive capacity (i.e. 19% higher than the baseline).

Another way to explain the fluctuation in LLM performance is the familiarity of the model with the games per se. We observe that more popular games (such as *Counter Strike*, *Apex Legends*, and *PUBG*, with a peak viewership<sup>10</sup> of 1,914,861, 674,070, and 597,663, respectively on Twitch<sup>11</sup>) yield generally better engagement predictions compared to less popular games (such as *HROT*, *Heretic*, and *Corridor 7*, with a peak viewership of 24,721, 2,280, and 195 on Twitch<sup>11</sup>), although we should be careful with naive over-generalizations

<sup>10</sup>Peak viewership refers to the historically highest number of concurrent viewers watching a stream. It is indicative of the maximum audience size.

<sup>11</sup>Numbers retrieved from <https://twitchtracker.com/>, January 2025.

from these findings. For one, *Counter Strike 1.6* is a variant of a very popular game with a peak viewership of 125,378 in itself, but the models struggle with correctly evaluating the change in engagement—at least in the *GameVibe* dataset. While it is possible that the training data of *GPT-4o* contains images from more popular games, attempting to verify this by reconstructing parts of the *GPT-4o* training data is out of scope of this paper.

## VI. DISCUSSION

The evaluation experiments presented in this paper are the first of its kind for LLM-based engagement prediction in games. While collectively we tried 2,440 combinations of experimental settings—varying the LLM model type, model size, prompting strategy, input type, and ground truth processing—there are still many aspects that we did not explore in this initial study. We argue, however, that we set out to lay ground works for future research by approaching the problem of automating gameplay annotation in a relatively straightforward way. While, for instance, we experimented with several out-of-box LLM models and prompting strategies, we kept the granularity of the vision input constant which potentially poses a core limitation to this initial study. Since we sample the videos in question at a 3-second interval, the model loses a lot of information between these frames. Although we briefly experimented with different time intervals (i.e. between 1 and 5 seconds), simply increasing the sampling rate did not yield a performance increase. It is likely, however, that by either providing more frames per query or using video input directly would lead to a significant performance improvement that remains to be tested in future studies. These investigations were purposefully left out of the scope of the current study, mainly because (at the time of writing) there were no widely available video models which could have fit into the experimental protocol presented here.

While video input could feed more information to the LLM, the context of the query could also be augmented, and then provided to the LLM, thereby improving its predictive capacity. By implementing a memory mechanism [67], for instance, we could potentially store and recall the temporal context of the play session, providing richer information to the model. Similarly, we could provide more context on the necessary domain knowledge for the task by implementing *retrieval-augmented generation* [68], where we could feed more information on the game, play session, or downstream task similarly to how we have been providing positive and negative examples to the model in our few-shot examples. We plan to pursue these avenues in our future studies in our effort to further investigate how more contextual information impacts the performance of LLMs towards fully autonomous engagement annotation.

Generating subjective labels is a relatively open field with a lot of unanswered questions. Naturally, the exploration should be extended into other datasets, involving games—also beyond first-person shooters—and other media as well. While this study focuses on engagement, there are other subjective aspects of both player and viewer experience that

could be evaluated further. A natural step forward would be to make use of a diverse set of affective corpora, focusing, for instance, on affect prediction across videogame datasets [69], but also architectural spaces [70] and movie corpora [71]. As discussed above, the current evaluation of LLMs—even though it was multimodal—considered a predetermined number of modalities: text and images. As we move forward and more multimodal architectures become widely adopted, the research into utilizing LLMs for autonomous affect annotation could encompass different modalities from images, through video, to audio. When it comes to interactive mediums such as games, user behavioural data could also be included [31] providing a richer context to the models.

## VII. CONCLUSION

This paper explored a novel application of LLMs for autonomously annotating the continuous experience of viewers when consuming videos of first-person shooter video games from the *GameVibe* corpus. We conducted an in-depth analysis comparing multiple foundation models, including Open-AI’s *GPT* and the *LLaVa* model families, and evaluated their performance across different input modalities (i.e. multimodal, text-based) and prompting strategies (i.e. one-shot, few-shot). Our findings confirm that model size and prompting strategy have a critical impact on model performance. The LLMs presented here demonstrate promising capabilities on certain game elicitors—although their overall performance only marginally surpasses the baseline. Perhaps unsurprisingly, the games where LLMs are successful predicting the continuous change in engagement are popular games with easy-to-read graphical styles and concise gameplay. The gap in performance on these games compared to more challenging elicitors shows that while LLMs have potential, there is still a long road ahead towards automated continuous affect labelling using these type of foundation models.

As LLMs continue to scale and evolve, we believe their ability to capture subjective experiences will drastically improve—especially when incorporating richer multimodal inputs such as video, audio, and physiological signals. The annotation capacities of such LLM-based foundation affect models extends well beyond the domain of games to video-based general affect modelling, and human-computer interaction at large. Based on the results presented here, we believe future work could leverage LLMs as flexible and scalable annotators in a wide variety of dynamic and real-world settings.

## APPENDIX

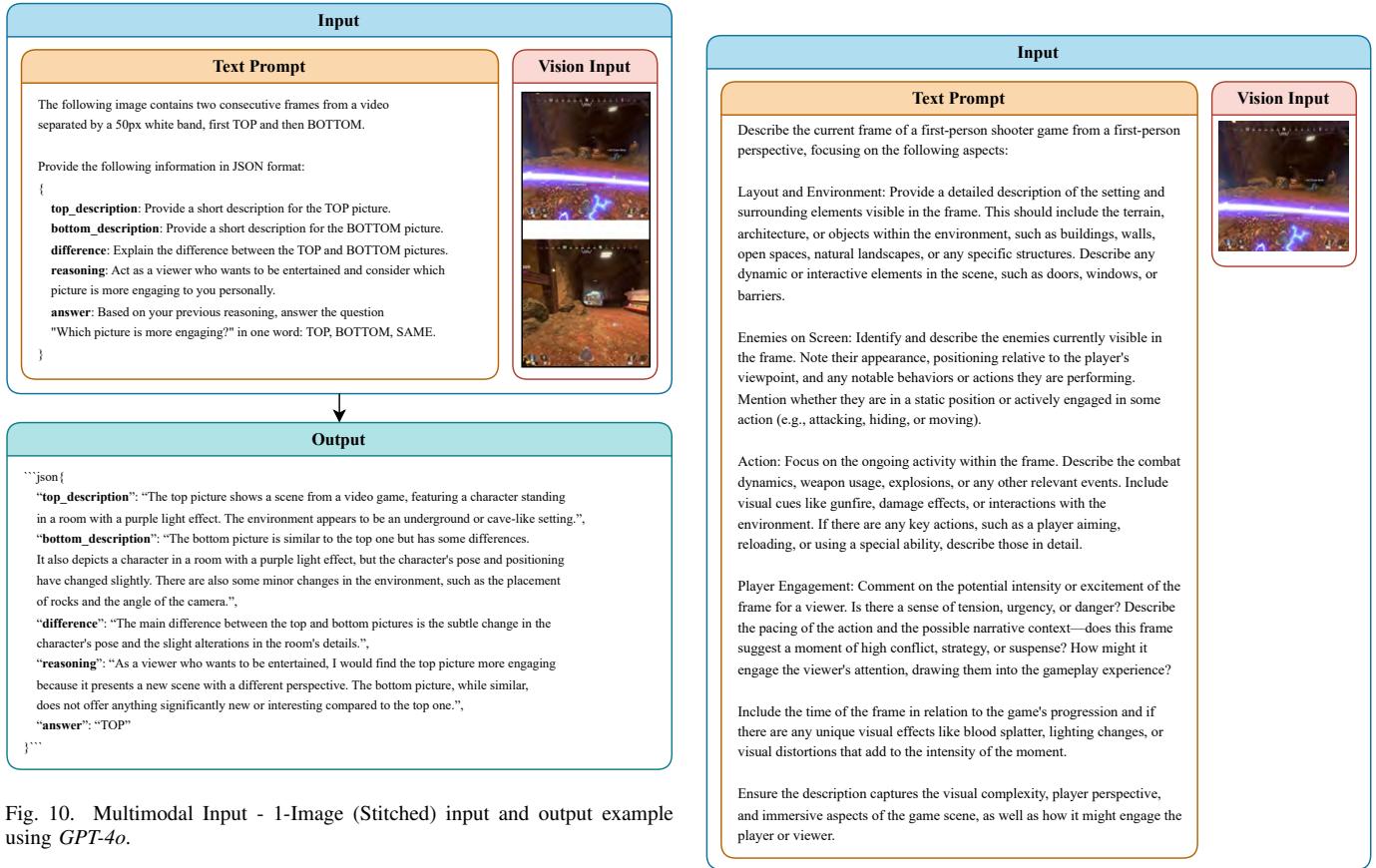


Fig. 10. Multimodal Input - 1-Image (Stitched) input and output example using *GPT-4o*.

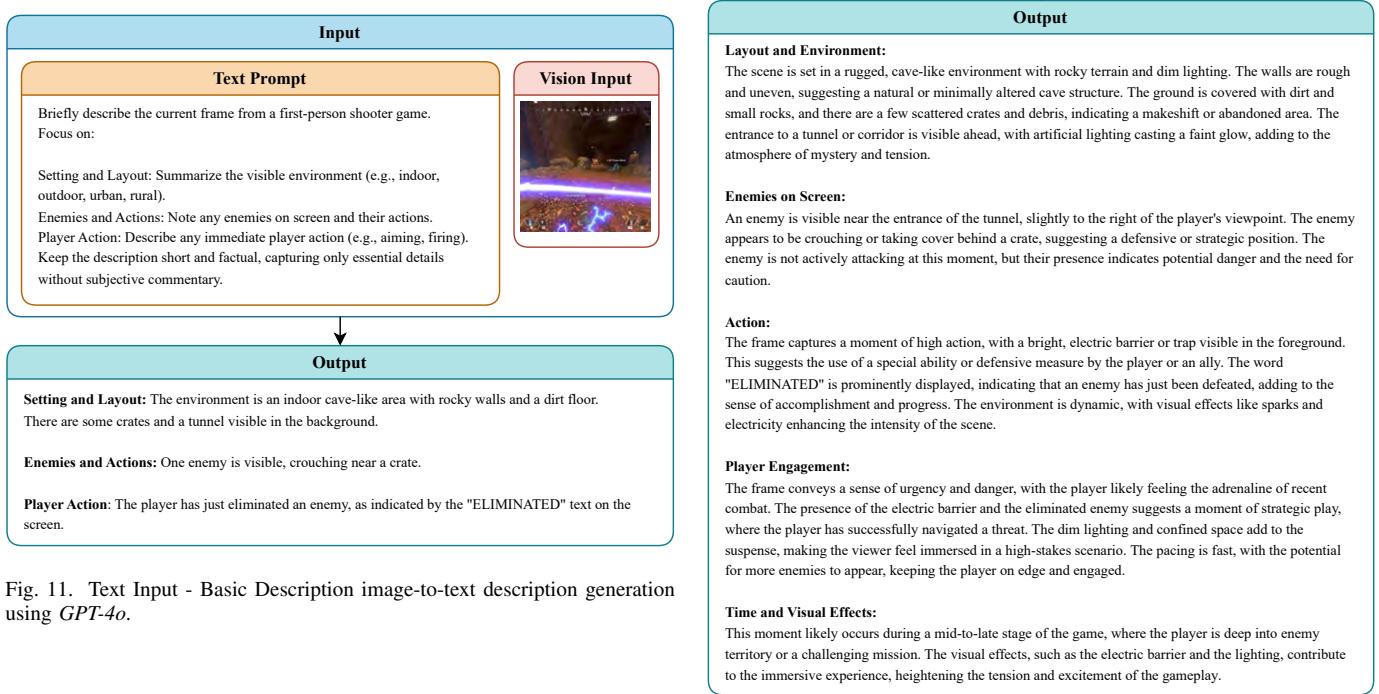
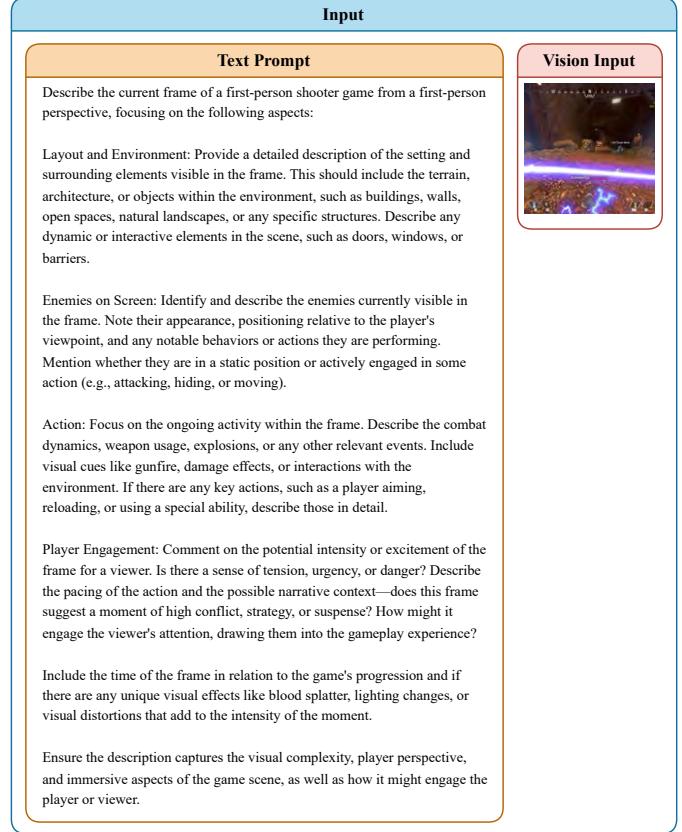


Fig. 11. Text Input - Basic Description image-to-text description generation using *GPT-4o*.

Fig. 12. Text Input - Advanced Description image-to-text description generation using *GPT-4o*.



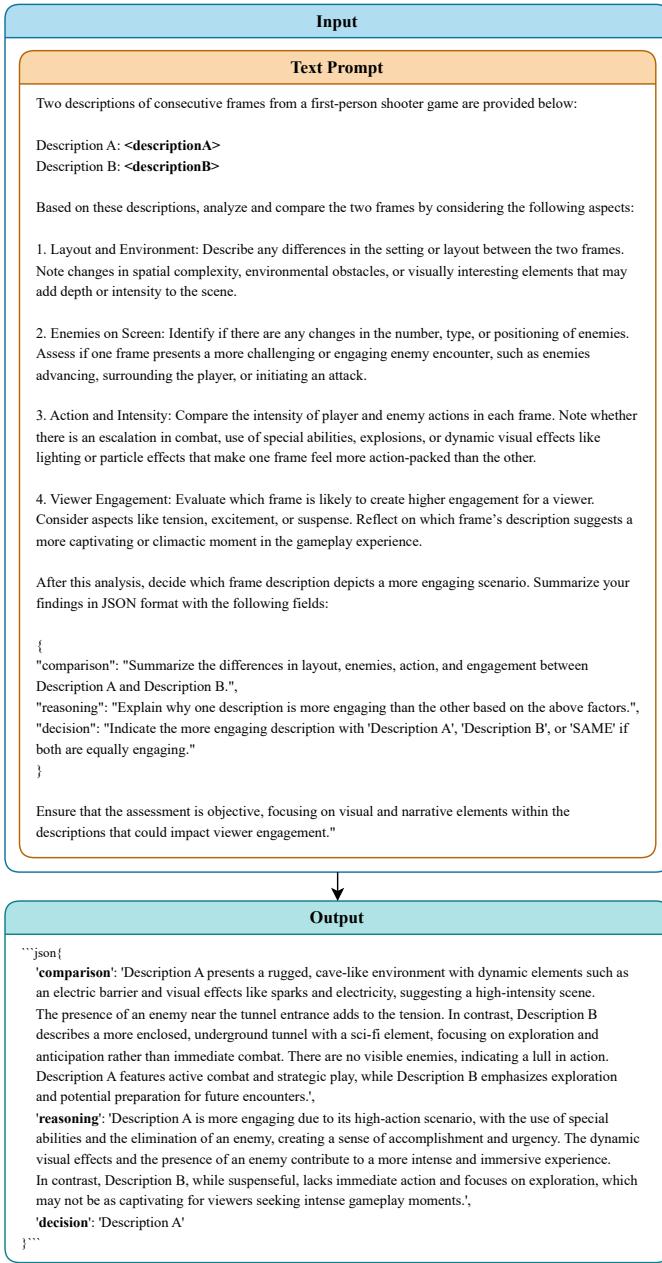


Fig. 13. Text Input - Engagement Prediction using *GPT-4o*. <description A> and <description B> are replaced with generated text descriptions.

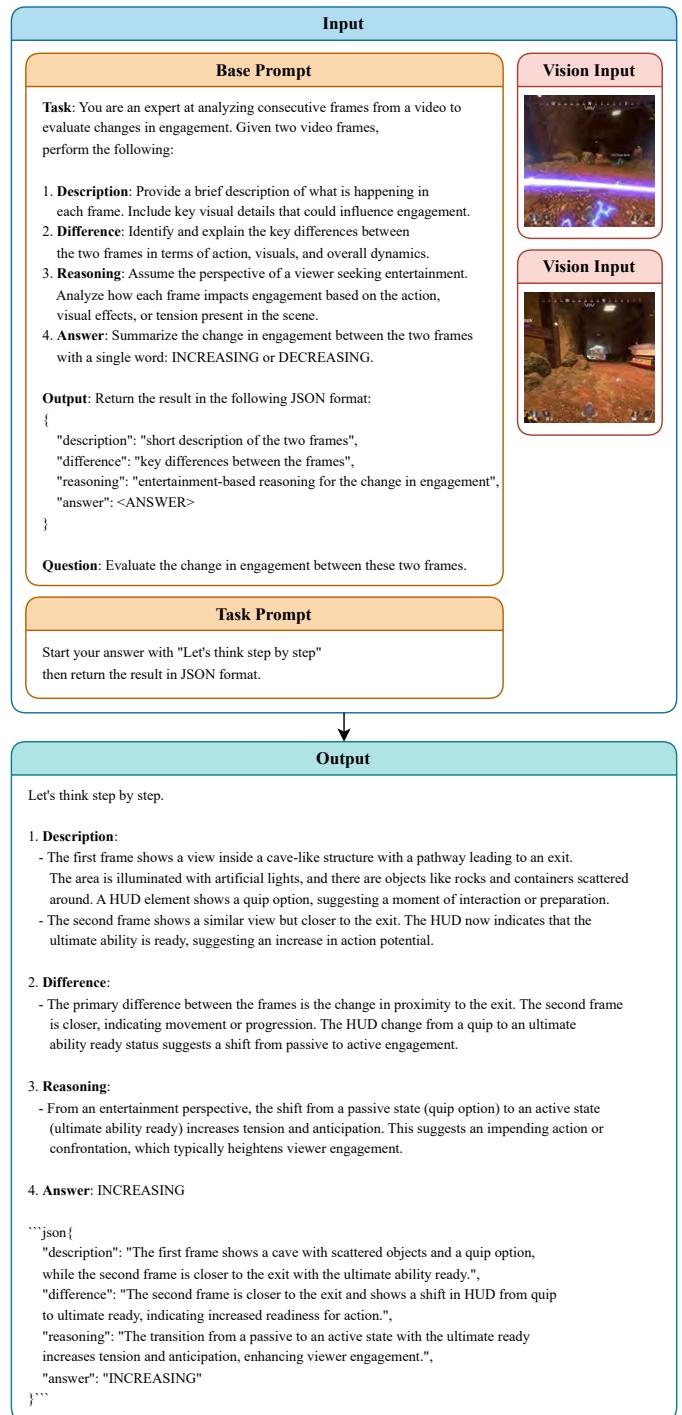


Fig. 14. Multimodal Input - 2-Image One-Shot input and output example using *GPT-4o*.

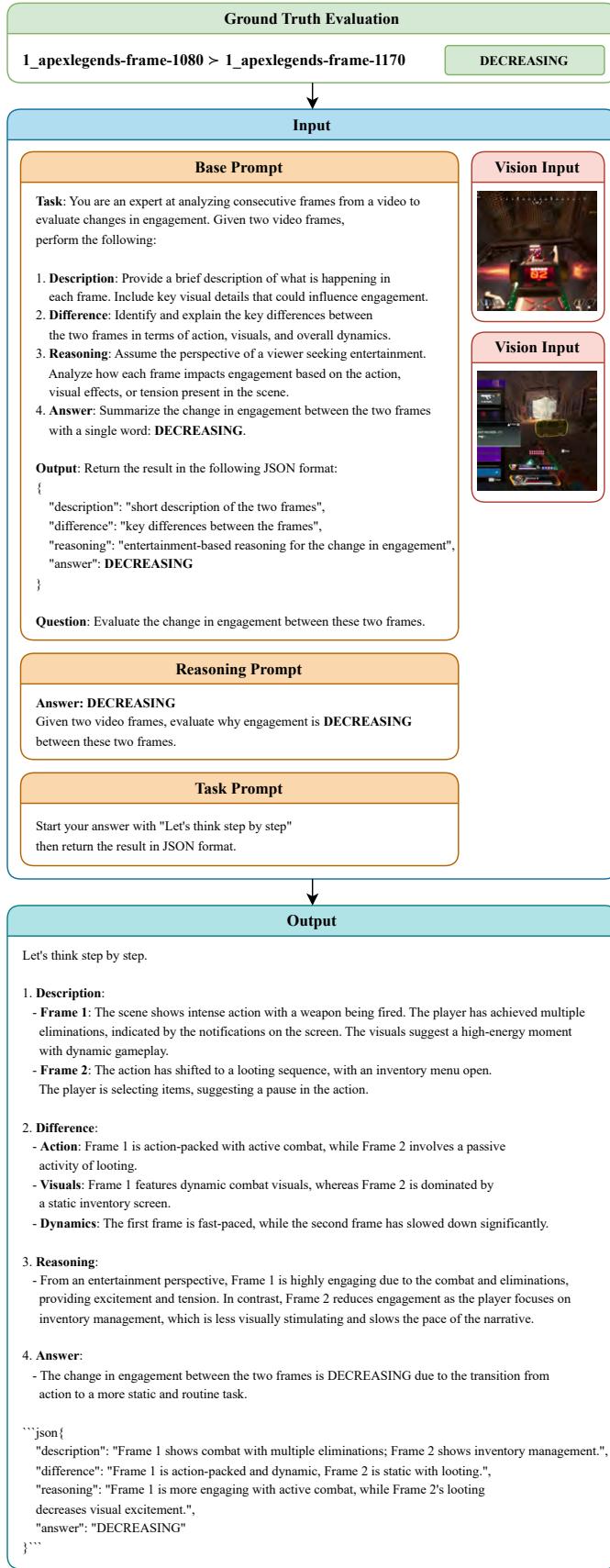


Fig. 15. Multimodal Input - 2-Image Few-Shot Example Reasoning Generation input and output example using *GPT-4o*. The generated output is used as an artificial example in the Few-Shot experiments.

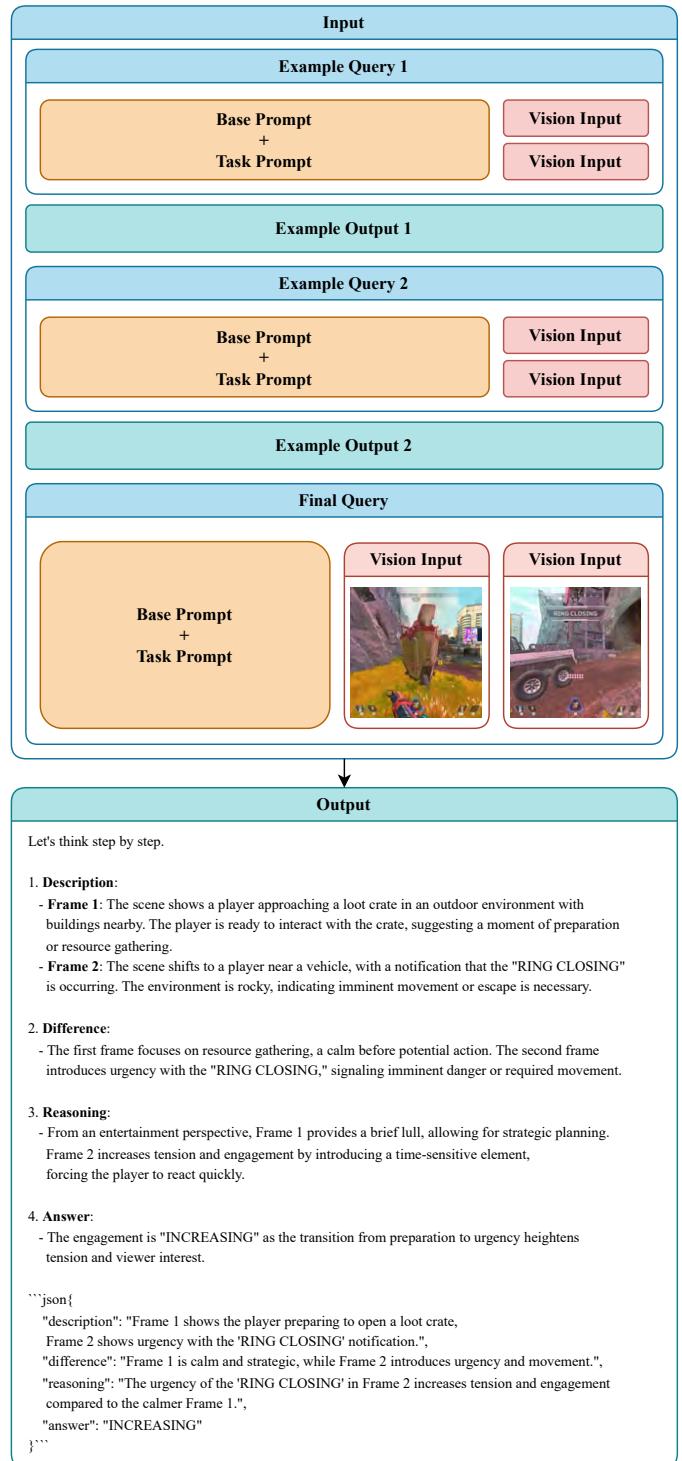


Fig. 16. Multimodal Input - 2-Image Few-Shot Engagement Prediction input and output example using *GPT-4o*. The Base and Task Prompts are identical to the ones described in Figure 14. Example Outputs are generated as shown in Figure 15.

## REFERENCES

- [1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [2] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing, “Sentiment analysis in the era of large language models: A reality check,” *arXiv preprint arXiv:2305.15005*, 2023.
- [3] J. Broekens, B. Hilpert, S. Verberne, K. Baraka, P. Gebhard, and A. Plaat, “Fine-grained affective processing capabilities emerging from large language models,” in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.
- [4] Y. Zhang, X. Yang, X. Xu, Z. Gao, Y. Huang, S. Mu, S. Feng, D. Wang, Y. Zhang, K. Song *et al.*, “Affective computing in the era of large language models: A survey from the nlp perspective,” *arXiv preprint arXiv:2408.04638*, 2024.
- [5] Z. Lian, L. Sun, H. Sun, K. Chen, Z. Wen, H. Gu, B. Liu, and J. Tao, “Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition,” *Information Fusion*, vol. 108, p. 102367, 2024.
- [6] X. Yang, W. Wu, S. Feng, M. Wang, D. Wang, Y. Li, Q. Sun, Y. Zhang, X. Fu, and S. Poria, “Mm-instructeval: Zero-shot evaluation of (multimodal) large language models on multimodal reasoning tasks,” *arXiv preprint arXiv:2405.07229*, 2024.
- [7] R. Gallotta, G. Todd, M. Zammitt, S. Earle, A. Liapis, J. Togelius, and G. N. Yannakakis, “Large language models and games: A survey and roadmap,” *arXiv preprint arXiv:2402.18659*, 2024.
- [8] G. N. Yannakakis and J. Togelius, *Artificial Intelligence and Games*. Springer, 2018.
- [9] M. Barthet, M. Kaselimi, K. Pinitas, K. Makantasis, A. Liapis, and G. N. Yannakakis, “Gamevibe: A multimodal affective game corpus,” *arXiv preprint arXiv:2407.12787*, 2024.
- [10] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, “The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection,” *IEEE transactions on affective computing*, vol. 14, no. 3, pp. 1743–1753, 2022.
- [11] M. M. Bradley and P. J. Lang, “Affective norms for english text (anet): Affective ratings of text and instruction manual,” *Techical Report. D-1, University of Florida, Gainesville, FL*, 2007.
- [12] P. Müller, A. Heimerl, S. M. Hossain, L. Siegel, J. Alexandersson, P. Gebhard, E. André, and T. Schneeberger, “Recognizing emotion regulation strategies from human behavior with large language models,” *arXiv preprint arXiv:2408.04420*, 2024.
- [13] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [14] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Huszenot, T. Mesnard, B. Shahriari, A. Ramé *et al.*, “Gemma 2: Improving open language models at a practical size,” *arXiv preprint arXiv:2408.00118*, 2024.
- [15] T. Schneeberger, M. Hladký, A.-K. Thurner, J. Volkert, A. Heimerl, T. Baur, E. André, and P. Gebhard, “The deep method: Towards computational modeling of the social emotion shame driven by theory, introspection, and social signals,” *IEEE Transactions on Affective Computing*, 2023.
- [16] G. Chochlakis, A. Potamianos, K. Lerman, and S. Narayanan, “The strong pull of prior knowledge in large language models and its impact on emotion recognition,” *arXiv preprint arXiv:2403.17125*, 2024.
- [17] S. Balloccu, P. Schmidtová, M. Lango, and O. Dušek, “Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms,” *arXiv preprint arXiv:2402.03927*, 2024.
- [18] J. Devlin, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [19] M. M. Amin, R. Mao, E. Cambria, and B. W. Schuller, “A wide evaluation of chatgpt on affective computing tasks,” *IEEE Transactions on Affective Computing*, 2024.
- [20] OpenAI, “GPT-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [21] J. Lai, W. Gan, J. Wu, Z. Qi, and S. Y. Philip, “Large language models in law: A survey,” *AI Open*, 2024.
- [22] Z. A. Nazi and W. Peng, “Large language models in healthcare and medical domain: A review,” in *Informatics*, vol. 11, no. 3. MDPI, 2024, p. 57.
- [23] S. Moore, R. Tong, A. Singh, Z. Liu, X. Hu, Y. Lu, J. Liang, C. Cao, H. Khosravi, P. Denny *et al.*, “Empowering education with llms-the next-gen interface and content generation,” in *International Conference on Artificial Intelligence in Education*. Springer, 2023, pp. 32–37.
- [24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. of the 31st Conference on Neural Information Processing Systems (NIPS)*, 2017.
- [25] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” *arXiv preprint arXiv:2001.08361*, 2020.
- [26] L. Floridi and M. Chiribatti, “Gpt-3: Its nature, scope, limits, and consequences,” *Minds and Machines*, vol. 30, no. 4, pp. 681–694, 2020.
- [27] C. F. Tsai, X. Zhou, S. S. Liu, J. Li, M. Yu, and H. Mei, “Can large language models play text games well? current state-of-the-art and open questions,” *arXiv preprint arXiv:2304.02868*, 2023.
- [28] S. Hu, T. Huang, F. Ilhan, S. Tekin, G. Liu, R. Komppela, and L. Liu, “A survey on large language model-based game agents,” *arXiv preprint arXiv:2404.02039*, 2024.
- [29] N. Ranella and M. Eger, “Towards automated video game commentary using generative ai.” in *EXAG@ AIIDE*, 2023.
- [30] T. Wang, M. Honari-Jahromi, S. Katsarou, O. Mikheeva, T. Panagiotakopoulos, S. Asadi, and O. Smirnov, “player2vec: A language modeling approach to understand player behavior in games,” *arXiv preprint arXiv:2404.04234*, 2024.
- [31] N. Rašajski, C. Trivedi, K. Makantasis, A. Liapis, and G. N. Yannakakis, “Behave: Behaviour alignment of video game encodings,” *arXiv e-prints*, pp. arXiv–2402, 2024.
- [32] X. You, P. Taveeikitworachai, S. Chen, M. Can Gursesli, X. Li, Y. Xia, and R. Thawomas, “Dungeons, dragons, and emotions: A preliminary study of player sentiment in llm-driven trpgs,” in *Proc. of the 19th International Conference on the Foundations of Digital Games*, 2024.
- [33] A. Zhu, L. Martin, A. Head, and C. Callison-Burch, “Calypso: Llms as dungeon master’s assistants,” in *Proc. of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 19, no. 1, 2023, pp. 380–390.
- [34] S. Sudhakaran, M. González-Duque, M. Freiberger, C. Glanois, E. Najarro, and S. Risi, “Mariogpt: Open-ended text2level generation through large language models,” in *Proc. of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 2024.
- [35] R. Gallotta, A. Liapis, and G. Yannakakis, “Llmaker: A game level design interface using (only) natural language,” in *Proc. of the IEEE Conference on Games (CoG)*, 2024.
- [36] D. Melhart, D. Gravina, and G. N. Yannakakis, “Moment-to-moment engagement prediction through the eyes of the observer: Pubg streaming on twitch,” in *Proc. of the Conference on the Foundations of Digital Games (FDG)*, 2020.
- [37] A. Canossa, D. Salimov, A. Azadvar, C. Harteveld, and G. Yannakakis, “For honor, for toxicity: Detecting toxic behavior through gameplay,” *Proc. of the ACM on Human-Computer Interaction*, vol. 5, no. CHI PLAY, pp. 1–29, 2021.
- [38] D. Melhart, A. Azadvar, A. Canossa, A. Liapis, and G. N. Yannakakis, “Your gameplay says it all: Modelling motivation in Tom Clancy’s The Division,” in *Proc. of the IEEE Conference on Games (CoG)*, 2019.
- [39] M. S. El-Nasr, A. Drachen, and A. Canossa, *Game analytics*. Springer, 2016.
- [40] S. Makarovych, A. Canossa, J. Togelius, and A. Drachen, “Like a DNA string: Sequence-based player profiling in Tom Clancy’s The Division,” in *Proc. of the Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*. York, 2018.
- [41] K. Makantasis, D. Melhart, A. Liapis, and G. N. Yannakakis, “Privileged information for modeling affect in the wild,” in *Proc. of the Conference on Affective Computing and Intelligent Interaction (ACII)*, 2021.
- [42] D. Melhart, A. Liapis, and G. N. Yannakakis, “Towards general models of player experience: A study within genres,” in *IEEE Conference on Games (CoG)*. IEEE, 2021, pp. 01–08.
- [43] B. M. Booth and S. S. Narayanan, “People make mistakes: Obtaining accurate ground truth from continuous annotations of subjective constructs,” *Behavior Research Methods*, vol. 56, no. 8, pp. 8784–8800, 2024.
- [44] G. N. Yannakakis, R. Cowie, and C. Busso, “The ordinal nature of emotions: An emerging approach,” *IEEE Transactions on Affective Computing*, 2018.
- [45] ———, “The ordinal nature of emotions,” in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 248–255.
- [46] H. Helson, “Current trends and issues in adaptation-level theory.” *American psychologist*, vol. 19, no. 1, p. 26, 1964.
- [47] R. L. Solomon and J. D. Corbit, “An opponent-process theory of motivation: I. temporal dynamics of affect.” *Psychological review*, vol. 81, no. 2, p. 119, 1974.

- [48] A. R. Damasio, “The somatic marker hypothesis and the possible functions of the prefrontal cortex,” *Philosophical Trans. of the Royal Society of London*, vol. 351, no. 1346, pp. 1413–1420, 1996.
- [49] G. N. Yannakakis and H. P. Martinez, “Grounding truth via ordinal annotation,” in *2015 international conference on affective computing and intelligent interaction (ACII)*. IEEE, 2015, pp. 574–580.
- [50] R. Lotfian and C. Busso, “Practical considerations on the use of preference learning for ranking emotional speech,” in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5205–5209.
- [51] D. Melhart, K. Sfikas, G. Giannakakis, and G. Y. A. Liapis, “A study on affect model validity: Nominal vs ordinal labels,” in *Workshop on Artificial Intelligence in Affective Computing*. PMLR, 2020, pp. 27–34.
- [52] D. Melhart, A. Liapis, and G. N. Yannakakis, “Pagan: Video affect annotation made easy,” in *Proc. of the 8th IEEE international conference on affective computing and intelligent interaction (ACII)*, 2019, pp. 130–136.
- [53] P. Lopes, G. N. Yannakakis, and A. Liapis, “Ranktrace: Relative and unbounded affect annotation,” in *Proc. of the 7th IEEE International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 158–163.
- [54] K. Pinitas, N. Rasajski, M. Barthet, M. Kaselimi, K. Makantasis, A. Liapis, and G. N. Yannakakis, “Varying the context to advance affect modelling: A study on game engagement prediction,” in *Proc. of the IEEE International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2024.
- [55] K. Pinitas, K. Makantasis, and G. N. Yannakakis, “Across-game engagement modelling via few-shot learning,” *arXiv preprint arXiv:2409.13002*, 2024.
- [56] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *Proc. of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1–19, 2024.
- [57] H. Liu, C. Li, Y. Li, and Y. J. Lee, “Improved baselines with visual instruction tuning,” 2023.
- [58] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [59] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, “Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality.” See <https://vicuna.lmsys.org> (accessed 14 April 2023), vol. 2, no. 3, p. 6, 2023.
- [60] OpenAI, “GPT-4o system card,” *arXiv preprint arXiv:2410.21276*, 2024.
- [61] A. Sergeyuk, Y. Golubev, T. Bryksin, and I. Ahmed, “Using ai-based coding assistants in practice: State of affairs, perceptions, and ways forward,” *Information and Software Technology*, vol. 178, p. 107610, 2025.
- [62] F. Meng, J. Wang, C. Li, Q. Lu, H. Tian, J. Liao, X. Zhu, J. Dai, Y. Qiao, P. Luo *et al.*, “Mmuu: Multimodal multi-image understanding for evaluating large vision-language models,” *arXiv preprint arXiv:2408.02718*, 2024.
- [63] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.
- [64] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, “Multimodal chain-of-thought reasoning in language models,” *arXiv preprint arXiv:2302.00923*, 2023.
- [65] A. Ajith, C. Pan, M. Xia, A. Deshpande, and K. Narasimhan, “Instructeval: Systematic evaluation of instruction selection methods,” *arXiv preprint arXiv:2307.00259*, 2023.
- [66] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, “Rethinking the role of demonstrations: What makes in-context learning work?” 2022. [Online]. Available: <https://arxiv.org/abs/2202.12837>
- [67] Z. Zhang, X. Bo, C. Ma, R. Li, X. Chen, Q. Dai, J. Zhu, Z. Dong, and J.-R. Wen, “A survey on the memory mechanism of large language model based agents,” *arXiv preprint arXiv:2404.13501*, 2024.
- [68] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive nlp tasks” in *Proc. of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, ser. NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., 2020.
- [69] D. Melhart, A. Liapis, and G. N. Yannakakis, “The arousal video game annotation (again) dataset,” *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2171–2184, 2022.
- [70] E. Xylakis, A. Liapis, and G. N. Yannakakis, “Affect in spatial navigation: A study of rooms,” *IEEE Transactions on Affective Computing*, pp. 1–11, 2024.
- [71] J. M. Girard, Y. Tie, and E. Liebenthal, “Dynamics: The dynamic affective movie clip database for subjectivity analysis,” in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.



**David Melhart** is a Postdoctoral Fellow at the Institute of Digital Games, University of Malta. He received his Ph.D. from the University of Malta, focusing on user research, player modeling, and affective computing in games. His research explores automated annotation, user modelling in videogames, and ethical AI. He has contributed to various academic and industry events, serving as Communication Chair (*FDG 2020, 2022; DiGRA 2025*), Workshop and Panels Chair (*FDG 2023*), and Workshop Organizer (*CHI-Play Workshop on Ethics and Transparency in Game Data 2024*). He is one of the main organizers of the *International Summer School on Artificial Intelligence and Games* (2018–2025). He has been a Review Editor of *Frontiers in Human-Media Interaction*, Guest Associate Editor of the *Frontiers in Virtual Reality and Human Behaviour* also serves as an *Editorial Assistant* for the *IEEE Transactions on Games*.



**Matthew Barthet** received a bachelors of science degree in computer science, and a masters of science degree in digital games from the University of Malta in 2019 and 2021, respectively. He is currently a PhD candidate at the University of Malta researching training reinforcement learning agents in affective computing applications. His other research interests include procedural content generation, game artificial intelligence, and computational creativity.



**Georgios N. Yannakakis** is a Professor and Director of the Institute of Digital Games, University of Malta (UM) and a co-founder of modl.ai. He received the PhD degree in Informatics from the University of Edinburgh in 2006. Prior to joining UM, in 2012 he was an Associate Professor at the Center for Computer Games Research at the IT University of Copenhagen. He does research at the crossroads of artificial intelligence, affective computing, games and computational creativity. He has published more than 300 papers in the aforementioned fields and his work has been cited broadly. His research has been supported by numerous national and European grants (including a Marie Skłodowska-Curie Fellowship) and has appeared in *Science Magazine* and *New Scientist* among other venues. He is currently the Editor in Chief of the *IEEE Transactions on Games*, an Associate Editor of the *IEEE Transactions on Evolutionary Computation*, and used to be Associate Editor of the *IEEE Transactions on Affective Computing* and the *IEEE Transactions on Computational Intelligence and AI in Games* journals. He has been the General Chair of key conferences in the area of game artificial intelligence (IEEE CIG 2010) and games research (FDG 2013, 2020). Among the several rewards he has received for his papers he is the recipient of the *IEEE Transactions on Affective Computing Most Influential Paper Award* and the *IEEE Transactions on Games Outstanding Paper Award*. Georgios is an IEEE Fellow.