Judging the Imitation Game: Exploring Creator Bias in VLM-Based PCG Level Evaluation

Yucheon Park, Geum-Hwan Hwang, Kyung-Joong Kim
Gwangju Institute of Science and Technology (GIST), South Korea
{yucheon6000, ghkdrmaghks}@gm.gist.ac.kr, kjkim@gist.ac.kr

***Abstract*—Quantitative metrics for Procedural Content Generation (PCG) evaluation struggle to capture subjective player experience and cognitive biases. This study examines whether Vision-Language Models (VLMs) can support human-like PCG level evaluation from gameplay videos, focusing on belief-conditioned and prompt-induced creator bias—the phenomenon where evaluations differ based on the perceived origin of the content. Using 60 gameplay videos, we evaluated five dimensions: fun, challenging, frustrating, surprising, and design. In the inferred-creator setting, the VLM showed evaluation trends consistent with human players in three of the five dimensions, with challenging and frustrating being the exceptions. These divergences suggest that some experiential dimensions, particularly those related to difficulty and negative affect, are not fully captured from gameplay videos alone. Importantly, the VLM's evaluations were shaped not by the actual generator category, but by its inferred belief about the creator. When creator labels were explicitly injected into the prompt, the VLM showed broad score inflation under the human-created condition, including frustrating. Despite these absolute score shifts, relative level rankings remained highly consistent. These results suggest that VLMs can approximate human-like subjective PCG evaluation, but creator information should be omitted from prompts for authorship-independent assessment.**

***Index Terms*—Procedural generation, Vision language model, Evaluation models, Cognition, Generative AI, Prompt engineering**

I. Introduction

Procedural Content Generation (PCG) has been widely used to reduce development costs and increase content diversity [1]. While prior PCG evaluation has often focused on functional and structural properties of generated levels, such metrics do not fully capture how players perceive level quality as an experience. In particular, player judgments can depend on dynamic and contextual factors that emerge through interaction with the content, rather than on static level structure alone [2], [3].

Gameplay videos provide a compact representation of both level structure and temporal interaction, including movement, failure, success, and progression. This makes Vision-Language Models (VLMs) a promising candidate for evaluating generated game levels through natural-language criteria. However, it remains unclear whether VLMs can approximate human-like evaluation trends, including subjective preferences and creator-related biases.

979-8-3315-9476-3/26/$31.00 ©2026 IEEE

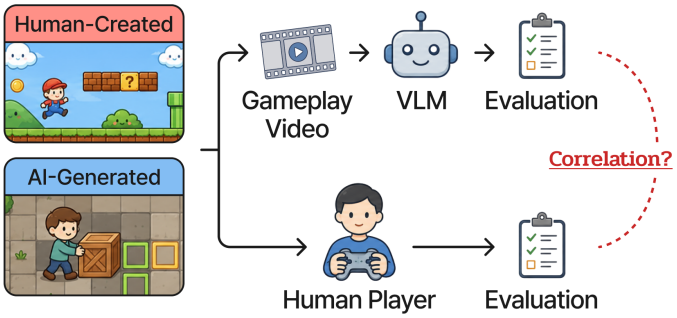


Fig. 1. Overview of the VLM-based game level evaluation settings used in this study.

External context can shape evaluation beyond the content itself. Prior work shows that judgments can differ depending on whether the same content is perceived as AI-generated [4]. This raises the need to examine both the evaluative potential of VLMs and their susceptibility to creator-related bias.

In this study, we analyze VLM-based PCG evaluation from two perspectives, following the evaluation settings shown in Fig. 1. First, we examine whether VLM evaluations based on gameplay videos show trends similar to those observed in prior human-subject studies. Second, we analyze how explicitly providing creator information in the evaluation prompt affects the VLM's judgment and experiential scores. Through this analysis, we investigate the potential of VLMs to reflect dynamic and context-dependent human evaluation, as well as the presence and impact of creator bias.

II. Background

A. PCG Evaluation

The evaluation of generated content is a core process in PCG research for validating both generation models and their outputs. In game level generation, prior studies have mainly relied on quantitative metrics that capture functional and structural properties, including playability, solvability, path structure, connectivity, and density [5]–[7]. These metrics are useful for verifying whether generated levels function properly and for comparing different generators, but they do not fully explain how players perceive level quality.

Structural validity does not necessarily correspond to player-perceived quality. A functionally valid level may not be experienced as interesting, appropriately challenging, or well-designed. Prior studies have therefore modeled player experience using level design features and player behavioral

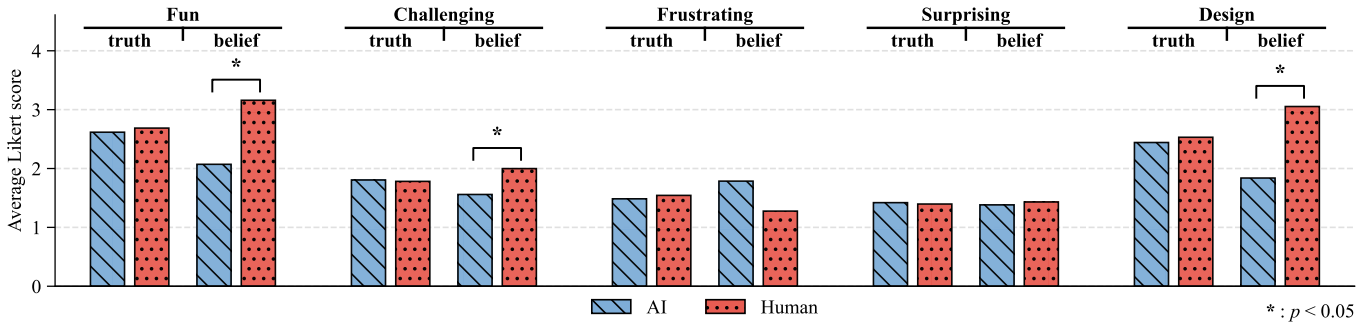


Fig. 2. Comparison of the average scores for each evaluation item in VLM-based game level evaluation. AI and Human indicate the generator type of the level, and the results are grouped by the actual generator label (truth) and the generator label inferred by the VLM (belief). The evaluation items consist of fun, challenging, frustrating, surprising, and design.

data [2], [3], or incorporated subjective factors such as enjoyment, aesthetics, and difficulty through human assessment [8]. Nevertheless, these approaches still largely rely on static level properties or post-play assessments, making it difficult to explain how players experience level elements over time and within interactive contexts.

B. LLM/VLM-as-a-Judge

With the emergence of large language models (LLMs), LLM-as-a-Judge has been studied as a way to evaluate generated outputs using natural-language criteria. It has been applied to tasks where fixed answers are difficult to define or human preference is central, such as summarization, dialogue, and question answering [9]. Studies such as G-Eval and MT-Bench have shown that strong LLM judges can approximate human evaluation in open-ended tasks [10], [11]. At the same time, LLM-based evaluation is not a complete replacement for human evaluation, as it can be affected by several biases and by design choices in prompts, criteria, scoring scales, and comparison formats [9], [11].

This judge paradigm has recently been extended to multimodal settings. VLMs can interpret visual inputs such as images or videos according to natural-language criteria, and have been used to evaluate visual outputs and multimodal responses [12], [13]. This suggests that VLM-as-a-Judge can provide a complementary approach to PCG evaluation by interpreting generated levels and gameplay videos in ways that go beyond conventional quantitative metrics, including their spatial, visual, temporal, and experiential qualities. However, it remains unclear how reliably VLMs evaluate generated game content and how their judgments are influenced by contextual information about the content, such as whether it is labeled as human-created or AI-generated.

III. System Implementation

a) Level Dataset: We utilized a total of 60 levels obtained from M. Bazzaz and S. Cooper [4], comprising 30 levels each for Super Mario Bros. and Sokoban (15 human-created and 15 AI-generated per game). The AI-generated levels include outputs based on large language models (LLMs), reinforcement learning (RL), and wave function collapse (WFC); detailed information about these AI-generated levels is provided in the appendix of the prior study [4].

b) Play Video Generation: Each level was played by AI agents (breadth-first search (BFS) and A*) within the Mario AI Framework [14] and PCG Benchmark [5] environments. The gameplay was recorded as mp4 videos—with resolutions of 256 × 256 px (00:02–01:07) for Super Mario Bros. and 144–192 × 128–192 px (00:04–01:44) for Sokoban, depending on the level—which served as the input for the VLM.

c) VLM-as-a-Judge: Drawing inspiration from G-Eval [10], we incorporated game-specific evaluation steps generated by Gemini 2.5 Pro into the VLM level evaluation prompts. To estimate token probabilities as described in G-Eval [10], we performed 20 independent inference calls per level using Gemini 2.5 Pro with Temperature 1.0 and Top-P 1.0. The final creator judgment was determined by the mode, and the evaluation scores were calculated as the mean of these 20 responses for the five dimensions evaluated in the prior study [4]: **fun**, **challenging**, **frustrating**, **surprising**, and **design**.

IV. Experimental Setup

To analyze the extent to which the VLM can emulate human-created level evaluation—including creator bias—and to assess its evaluation consistency, we conducted two experiments. Within each experiment, identical evaluation steps were applied per game, and the exact same gameplay videos were provided as input for each respective level. The source code and full evaluation prompts are publicly available in the repository[1].

A. Inferred-Creator Evaluation

For comparison with the prior study [4], the VLM was instructed to sequentially generate the following contents and return them in a JSON format:

1) A description explaining what occurred in the video.
2) The reasoning behind the creator judgment.
3) The creator selection (Human or AI).
4) The confidence level in the creator selection (on a 3-point scale).
5) 5-point Likert scale scores for five dimensions (fun, challenging, frustrating, surprising, design) regarding the level, accompanied by brief justifications.

[1]https://github.com/GIST-CILab/vlm-pcg-eval

TABLE I
COMPARISON OF MEANS, SCORE SHIFTS (Δ), AND BIAS SIGNIFICANCE BASED ON CREATOR BELIEF

Experience Metric	Original Paper (Human Players) [4]			Ours (VLM)			Bias Alignment
	AI	Human	Shift (Δ_O)	AI	Human	Shift (Δ_V)	
Enjoyment / Fun	2.92	3.72	**+0.80** ($p < 0.001$)	2.07	3.16	**+1.09** ($p < 0.001$)	**Match** (AI < Human)
Difficulty / Challenge	3.88	3.65	**-0.23** ($p < 0.015$)	1.56	2.00	**+0.44** ($p < 0.015$)	Mismatch
Frustration / Negative Affect	3.60	2.84	**-0.76** ($p < 0.001$)	1.79	1.28	-0.51 ($p = 0.263$)	Mismatch
Novelty / Surprise	2.65	2.62	-0.03 ($p = 0.96$)	1.38	1.43	+0.05 ($p = 0.48$)	**Match** (Both Insig.)
Aesthetics / Design Quality	2.70	3.57	**+0.87** ($p < 0.001$)	1.84	3.05	**+1.21** ($p < 0.001$)	**Match** (AI < Human)

Note: Shift (Δ) is calculated as (Belief: Human) $-$ (Belief: AI); Human player p-values were calculated using ordinal logistic regression with random intercepts for participant and unique game level [4], while VLM p-values were computed using two-sided Mann-Whitney U tests. 'AI < Human' indicates a significant positive shift (favoring Human), while 'AI > Human' indicates a significant negative shift (favoring AI). 'Both Insig.' denotes that neither human players nor the VLM showed a statistically significant bias based on belief ($p > 0.05$).

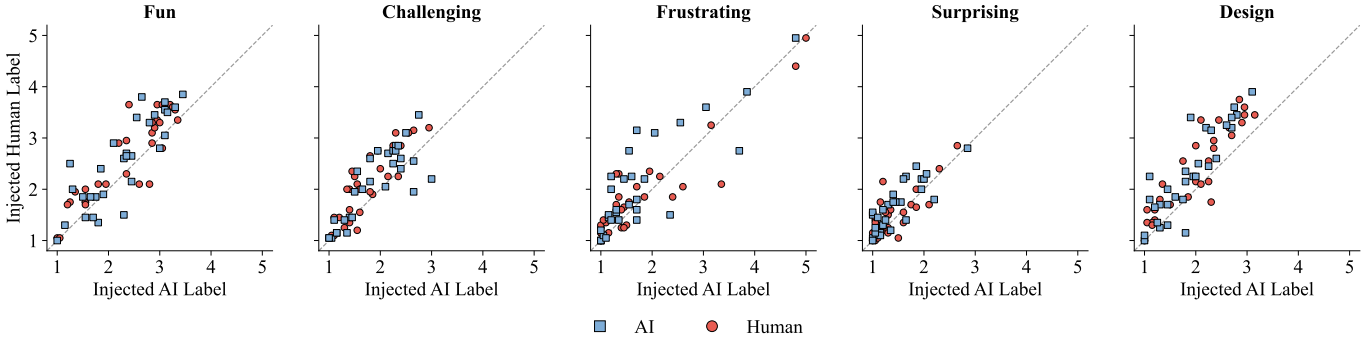


Fig. 3. Scatter plots of VLM evaluation scores for the same levels under two forced-label conditions: forced as AI-generated (x-axis) and forced as human-created (y-axis). Each point represents the pair of scores assigned to the same level when its presented creator label is changed.

B. Label-Injected Evaluation

Instead of requesting the VLM to deduce the creator, the evaluation was conducted by explicitly providing creator information within the input prompt. Consequently, a level actually created by a human could be presented as AI-generated, and vice versa. In this experiment, a cross-evaluation was performed on all 60 levels by independently injecting both the "human-created" condition and the "AI-generated" condition for every level.

V. RESULTS

A. Accuracy and Creator Bias

In *Inferred-Creator Evaluation*, the VLM's creator identification accuracy was 53.3%, which is close to chance level ($p = 0.699$, binomial test) and comparable to the 53% accuracy reported in the prior human-subject study involving 154 human players [4]. These results indicate that the VLM is unable to reliably determine whether a level is AI-generated.

When grouping the results of *Inferred-Creator Evaluation* by the actual creator, a Mann-Whitney U test (two-sided) showed no significant differences across any metrics ($p \geq 0.609$). In contrast, evaluations grouped by the VLM's belief exhibited statistically significant variations (Fig. 2). The VLM assigned higher scores for fun and design ($p < 0.001$), as well as challenging ($p < 0.015$), when it perceived a level as human-created. However, neither surprising ($p = 0.48$) nor frustrating ($p = 0.263$) showed a significant belief-based shift.

As summarized in Table I, the direction of the VLM's belief-based score shifts matched the prior human-subject study in three of the five dimensions: fun, surprising, and design. This suggests that the VLM's belief-conditioned score changes were not arbitrary, but partly reflected human-like bias directions. However, challenging and frustrating diverged from the human-subject pattern, indicating that the VLM may interpret dimensions related to difficulty and negative affect differently from human players. Overall, these results suggest that creator bias was driven not by the objective generator category, but by the VLM's inferred belief about the creator.

In *Label-Injected Evaluation*, where a uniform creator label was provided for all levels, a Mann-Whitney U test (two-sided) revealed that the actual creator again had no significant effect ($0.104 \leq p \leq 0.629$). However, based on the creator label provided in the prompt, a Wilcoxon signed-rank test (two-sided) showed that the VLM evaluated the human-created condition significantly higher across all evaluation dimensions: fun, challenging, surprising, and design ($p < 0.001$), as well as frustrating ($p = 0.001$). This label-injected score shift is further illustrated in Fig. 3, where each point compares the VLM's scores for the same level under the AI-labeled and human-labeled conditions. While *Inferred-Creator Evaluation* demonstrates an intrinsic bias driven by visual cues, *Label-Injected Evaluation* confirms that explicit information in the prompt induces a substantial bias in the VLM.

B. Evaluation Consistency

Despite the presence of these biases, the VLM's relative evaluations maintained consistency. Measuring Kendall's Tau-b (two-sided test)—which is robust to tied ranks for the

averaged 5-point scale scores—revealed high rank correlations across all metrics between the conflicting injected conditions in *Label-Injected Evaluation* ($0.586 \leq \tau \leq 0.731$, $p < 0.001$). This indicates that although injected creator labels shifted the absolute scores, the VLM preserved a relatively stable ordering of levels within each metric.

VI. Discussion and Conclusion

Through this study, we observed that a Vision-Language Model (VLM) can produce evaluation trends partially consistent with human players in game level evaluation, while also exhibiting creator-related bias. In the inferred-creator setting, the VLM's evaluations were not determined by the actual creator of a level, but varied according to the creator label inferred by the model. This suggests that VLM-based evaluation reflects not only visual evidence from gameplay videos, but also belief-based assumptions about the content's origin. In the label-injected setting, explicitly providing creator information further shifted the VLM's scores, demonstrating that prompt context can substantially affect absolute evaluations.

Despite these shifts, the VLM maintained relatively stable rank correlations across conflicting injected-label conditions. This indicates that despite absolute score biases, the model preserves useful relative ordering. Therefore, VLM-as-a-Judge should be understood not as a neutral replacement for human evaluation, but as a complementary tool whose outputs depend on both visual input and prompt context. In particular, creator information should be omitted from prompts when the goal is to evaluate level quality independently of perceived authorship.

Distinct patterns were observed in the challenging and frustrating metrics, where the VLM's evaluations diverged from the directions observed in human players. In the prior study [4], participants exhibited a negative affective bias, associating perceived AI-generated levels with higher difficulty and frustration. However, the VLM did not show a significant belief-based shift in frustrating, which may reflect a form of *affective decoupling*: unlike human players, it does not directly experience frustration during gameplay. In contrast, the VLM evaluated perceived human-created levels as more challenging, suggesting that it may associate human design with deliberate structural complexity, whereas human players may perceive erratic or unoptimized AI-generated layouts as difficult.

Additionally, this challenge discrepancy may be attributed to the behavioral characteristics of the AI agents used to generate the gameplay videos. These agents often follow optimized paths and clear levels mechanically, thereby failing to capture mistakes, hesitation, repeated attempts, and other behavioral cues that shape human perceptions of challenge. In cases where the agent fails to complete a level, the VLM may also be forced to evaluate the level from incomplete or fragmented visual evidence.

Future work should therefore explore gameplay videos generated by agents with varying skill levels and behavioral styles, ranging from novice-like to expert-like play. Such videos may better represent the diversity of human player experiences and improve the evaluation of metrics that depend strongly on player behavior, such as challenge. More broadly, future studies should examine whether these findings generalize across different VLMs, game genres, and prompt designs. Overall, this study provides empirical evidence that VLMs can support qualitative PCG level evaluation, while highlighting the need for blind evaluation protocols and careful control of creator-related context.

Acknowledgment

Generative AI tools (e.g., Gemini) were used for code & image generation, translation, etc., with the authors assuming full responsibility for the final manuscript. This research was supported by the 'Project for science and technology opens the future of the region' program, funded by the Ministry of Science and ICT(MSIT) and Gwangju Metropolitan City, Republic of Korea in 2026 (Project Name:Convergence culture virtual studio for realizing artificial intelligence based metaverse).

References

[1] N. Shaker, J. Togelius, and M. J. Nelson, "Procedural content generation in games," 2016.

[2] C. Pedersen, J. Togelius, and G. N. Yannakakis, "Modeling player experience in super mario bros," in *2009 IEEE Symposium on Computational Intelligence and Games*. IEEE, 2009, pp. 132–139.

[3] ——, "Modeling player experience for content creation," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 2, no. 1, pp. 54–67, 2010.

[4] M. Bazzaz and S. Cooper, "Playing the imitation game: How perceived generated content shapes player experience," in *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, ser. CHI '26. New York, NY, USA: Association for Computing Machinery, 2026. [Online]. Available: https://doi.org/10.1145/3772318.3790473

[5] A. Khalifa, R. Gallotta, M. Barthet, A. Liapis, J. Togelius, and G. N. Yannakakis, "The procedural content generation benchmark: an open-source testbed for generative challenges in games," in *Proceedings of the 20th International Conference on the Foundations of Digital Games*, 2025, pp. 1–12.

[6] G. Smith and J. Whitehead, "Analyzing the expressive range of a level generator," in *Proceedings of the 2010 workshop on procedural content generation in games*, 2010, pp. 1–7.

[7] O. Withington, M. Cook, and L. Tokarchuk, "On the evaluation of procedural level generation systems," in *Proceedings of the 19th International Conference on the Foundations of Digital Games*, 2024, pp. 1–10.

[8] L. H. Lelis, W. M. Reis, and Y. Gal, "Procedural generation of game maps with human-in-the-loop algorithms," *IEEE Transactions on Games*, vol. 10, no. 3, pp. 271–280, 2017.

[9] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu *et al.*, "A survey on llm-as-a-judge," *The Innovation*, 2024.

[10] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: NLG evaluation using gpt-4 with better human alignment," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2511–2522. [Online]. Available: https://aclanthology.org/2023.emnlp-main.153/

[11] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in neural information processing systems*, vol. 36, pp. 46 595–46 623, 2023.

[12] S. Lee, S. Kim, S. Park, G. Kim, and M. Seo, "Prometheus-vision: Vision-language model as a judge for fine-grained evaluation," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 11 286–11 315.

[13] D. Chen, R. Chen, S. Zhang, Y. Wang, Y. Liu, H. Zhou, Q. Zhang, Y. Wan, P. Zhou, and L. Sun, "Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark," in *Forty-first International Conference on Machine Learning*, 2024.

[14] amidos2006, "Mario ai framework." [Online]. Available: https://github.com/amidos2006/Mario-AI-Framework