

Judging the Imitation Game:

Exploring Creator Bias in VLM-Based PCG Level Evaluation

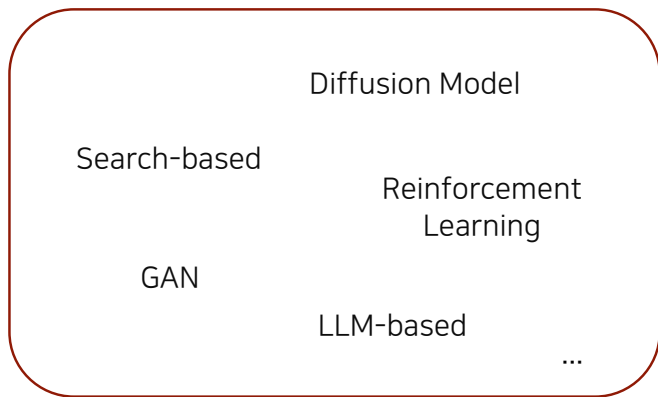
Yucheon Park, Geum-Hwan Hwang, Kyung-Joong Kim

{yucheon6000, ghkdrmaghks}@gm.gist.ac.kr, kjkim@gist.ac.kr

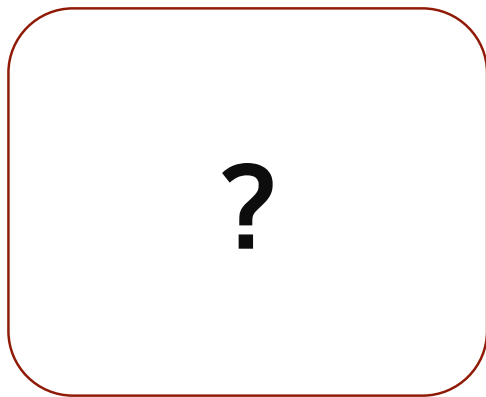
GIST, South Korea



Generating Content Is Getting Easier—Evaluating It Is Not



Procedural Content Generate



Evaluate

Quantitative / Structural Evaluation



Human Evaluation



Quantitative / Structural Evaluation



Human Evaluation



Subjective Experience Missing!

Quantitative / Structural Evaluation

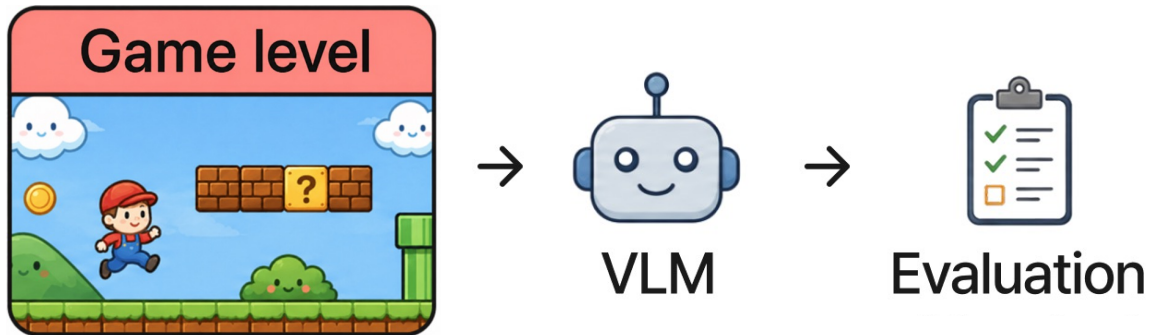


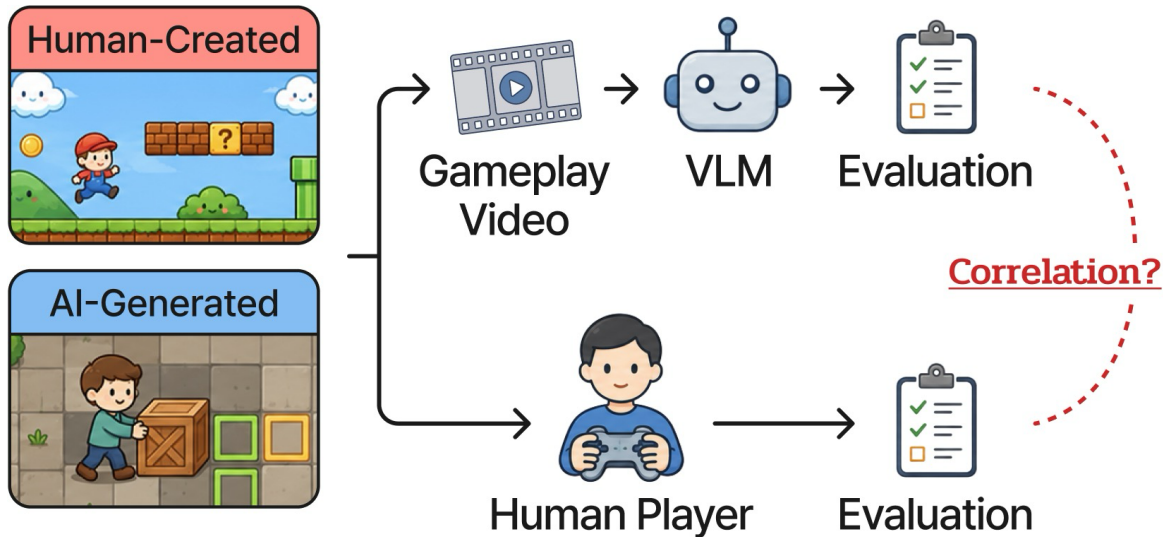
Subjective Experience Missing!

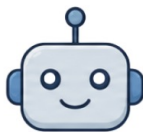
Human Evaluation



High Evaluation Cost!







VLM



Evaluation
x20 per level

① Creator Prediction

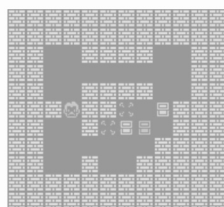
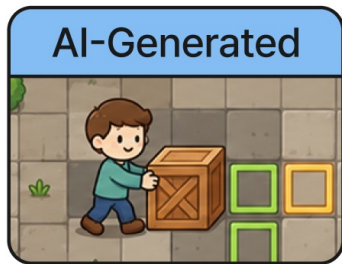
Human vs. AI

+

② Level Evaluation

Fun, Challenging,
Frustrating, Surprising,
Design

Gameplay
Video



Gameplay
Video



VLM



Evaluation
x20 per level

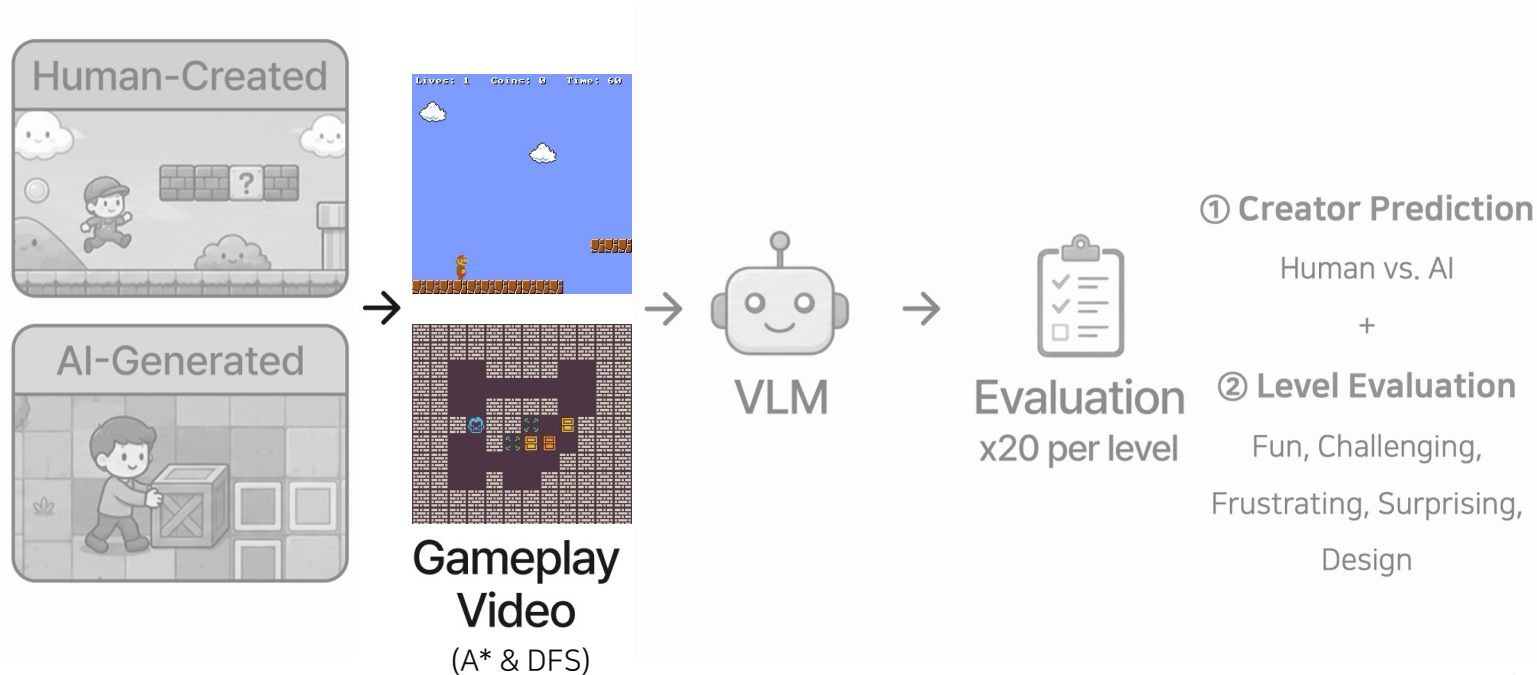
① Creator Prediction

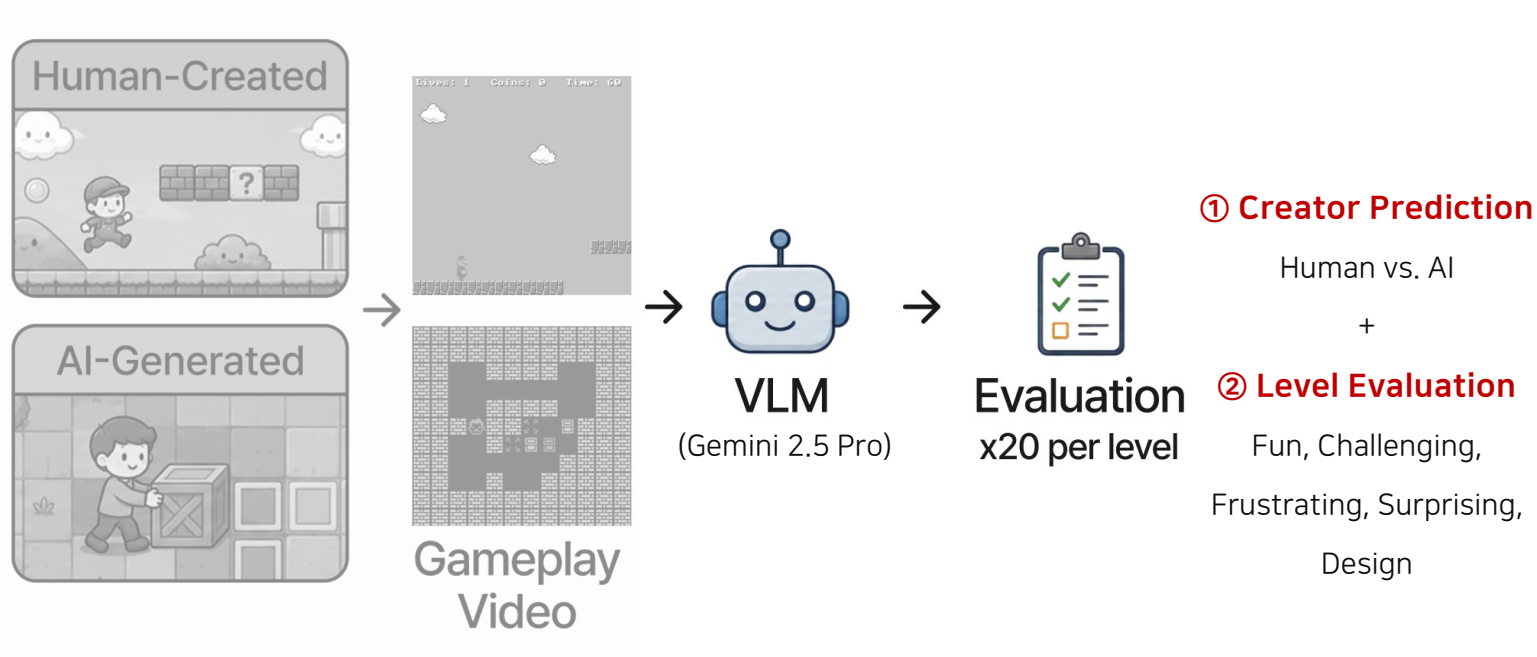
Human vs. AI

+

② Level Evaluation

Fun, Challenging,
Frustrating, Surprising,
Design





53.3%

VLM Accuracy

Both humans and VLMs
fail to identify the creator!

53.3%

VLM Accuracy

$\approx$

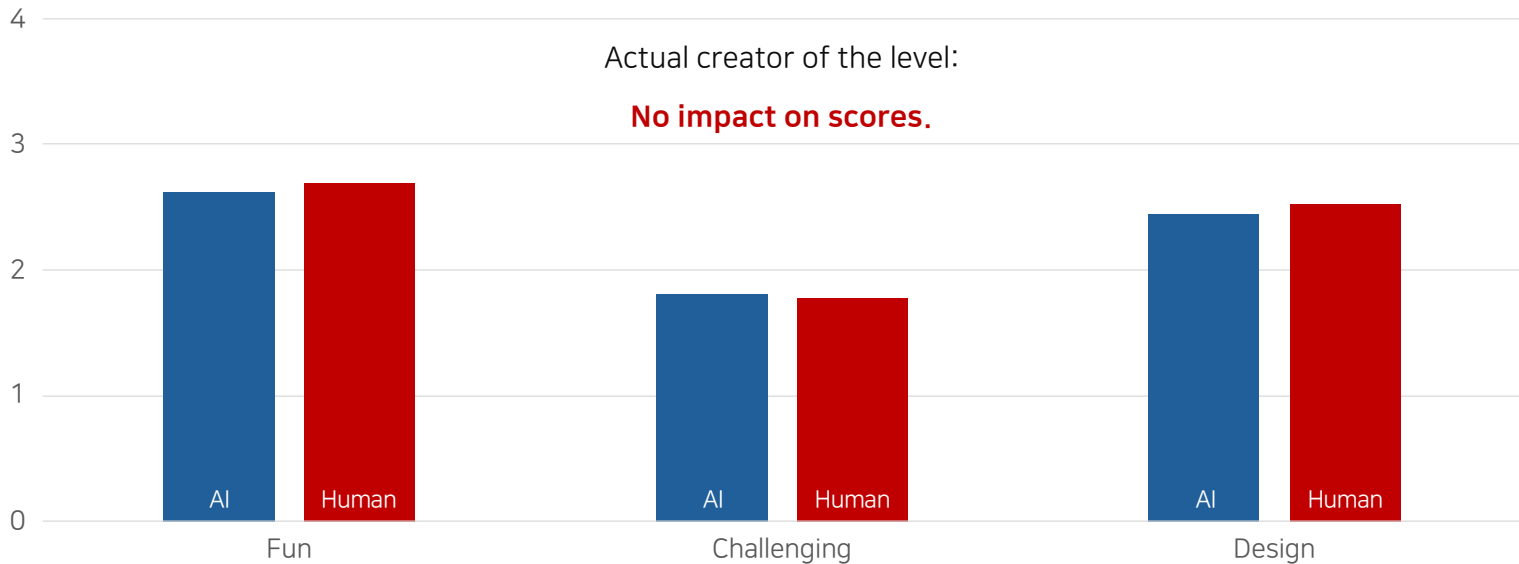
53%

Human Accuracy

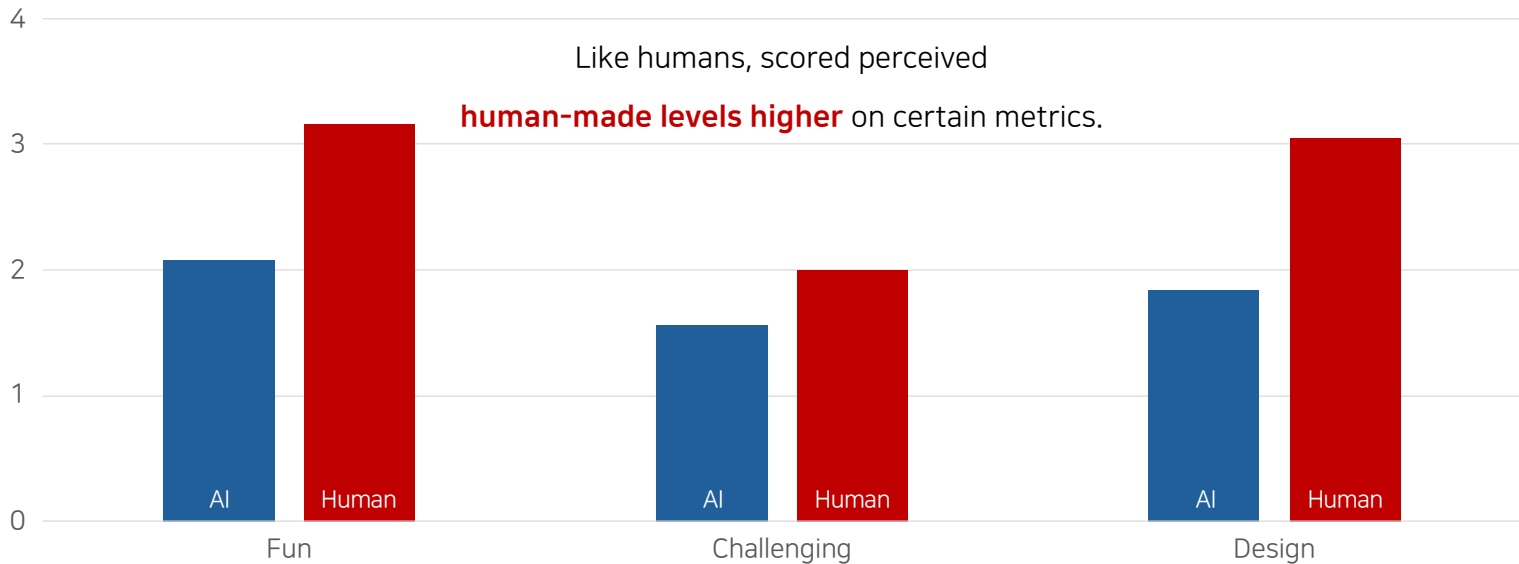
Ground Truth

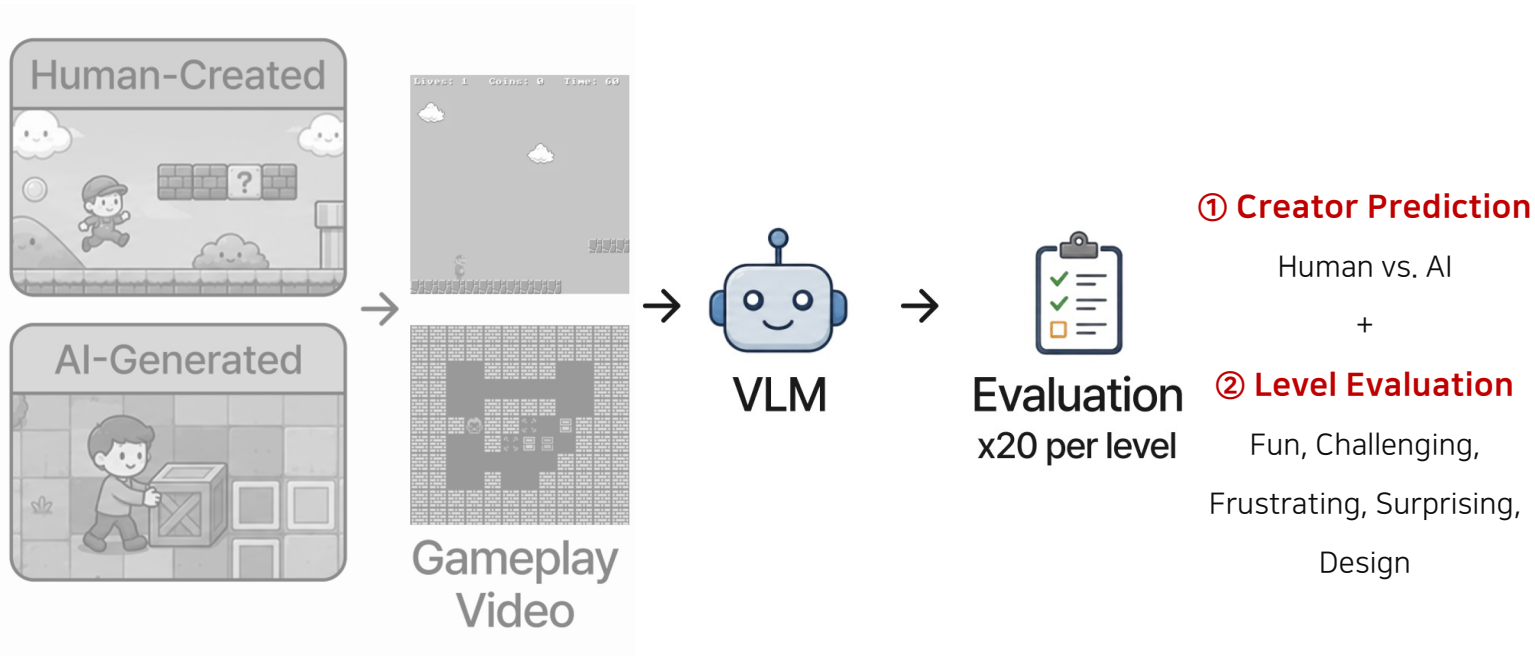
Actual creator of the level:

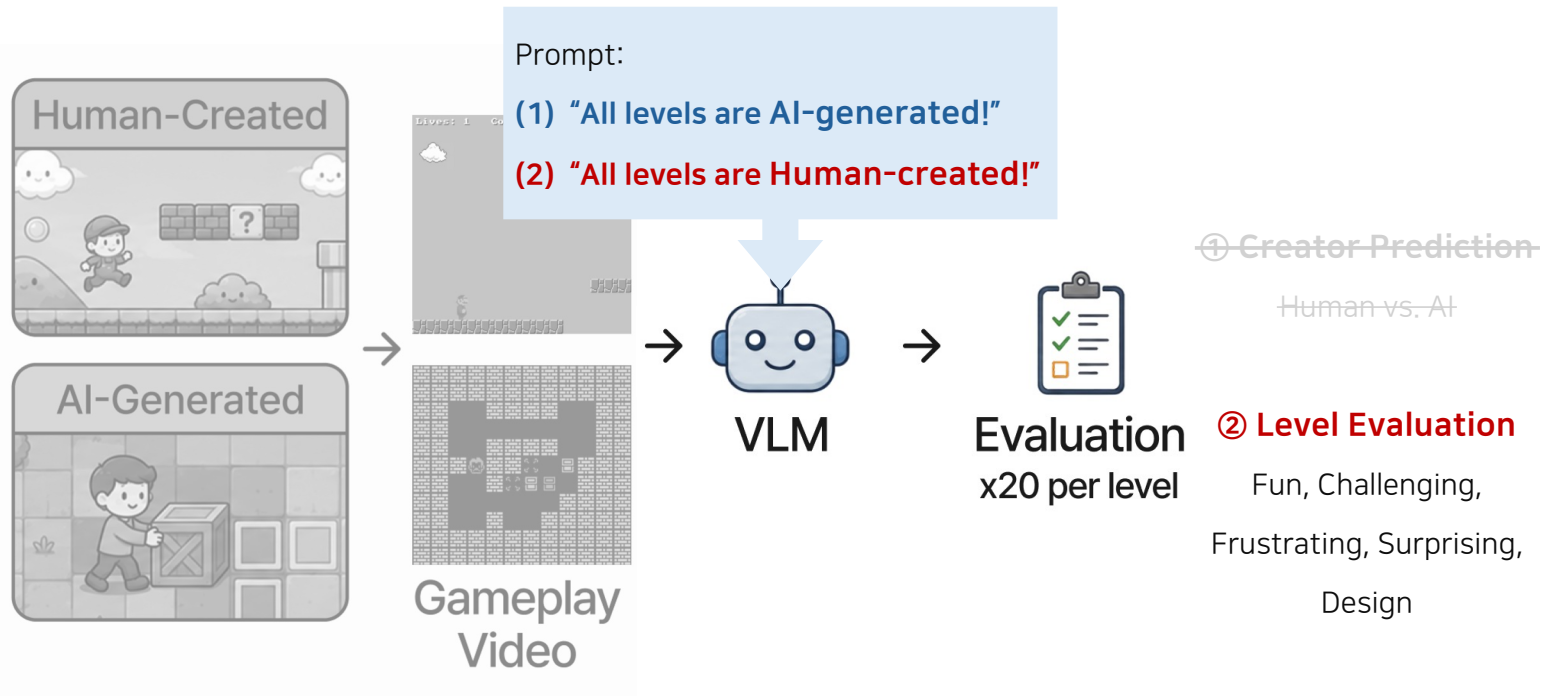
No impact on scores.



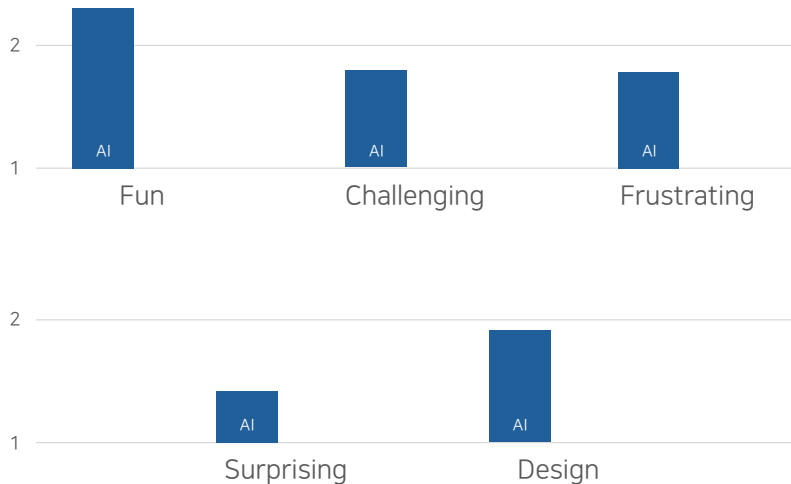
VLM Prediction







"All levels are AI-generated!"



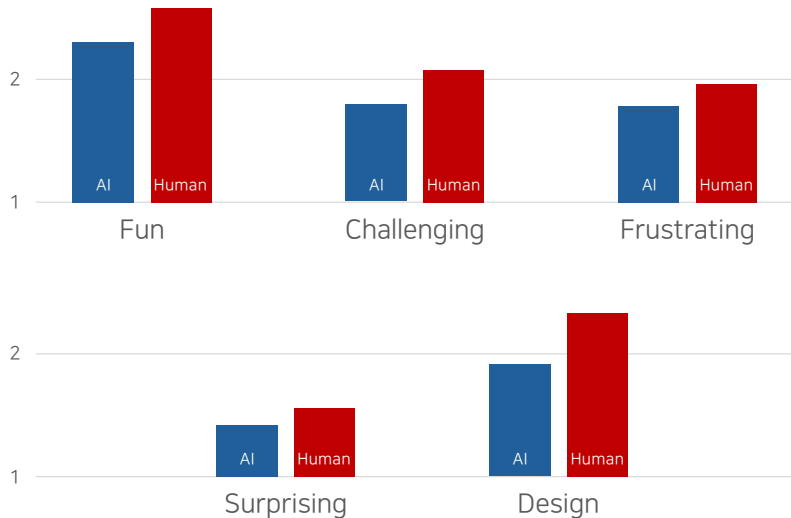
► Score Inflation

Stating "Human-Created" induced a systematic score increase across all metrics for identical levels.

► Consistent Relative Rankings

Although absolute scores shifted due to bias, the relative ordering of levels remained highly consistent across all evaluation metrics.

"All levels are Human-created!"



► Score Inflation

Stating "Human-Created" induced a systematic score increase across all metrics for identical levels.

► Consistent Relative Rankings

Although absolute scores shifted due to bias, the relative ordering of levels remained highly consistent across all evaluation metrics.

► **Creator Bias in VLMs**

Similar to humans, VLM-based evaluations exhibit a strong bias where absolute scores are significantly influenced by the inferred or explicitly prompted creator identity (Human vs. AI).

► **Evaluation Consistency and the Need for Blind Evaluation**

The relative rankings of levels remained highly consistent across all metrics.

Therefore, by applying a "blind evaluation" protocol that strictly omits creator information from prompts, VLMs demonstrate strong potential as a qualitative PCG evaluation tool.

Thank you!

Judging the Imitation Game:

Exploring Creator Bias in VLM-Based PCG Level Evaluation

Yucheon Park, Geum-Hwan Hwang, Kyung-Joong Kim

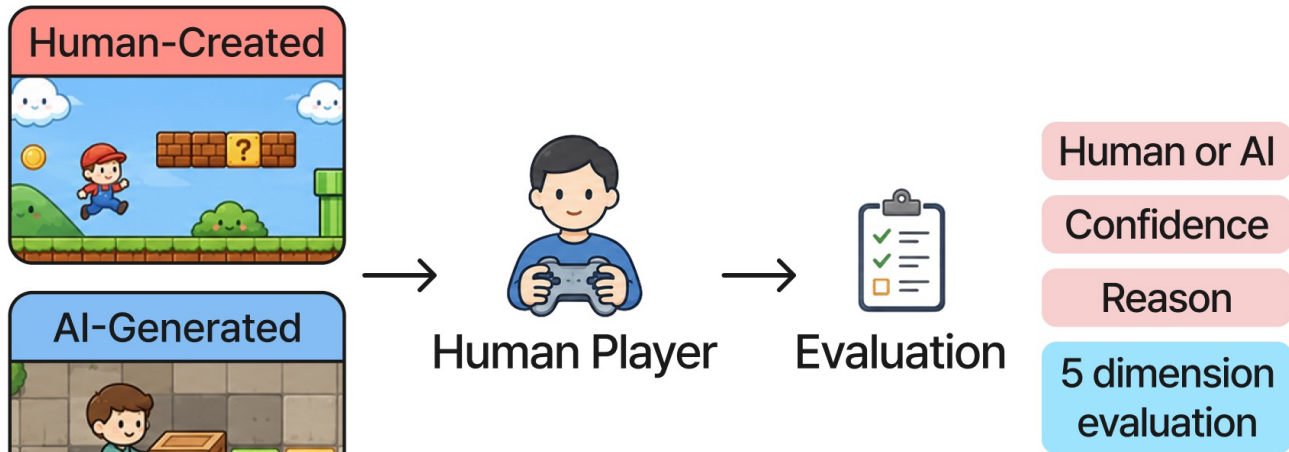
{yucheon6000, ghkdrmaghks}@gm.gist.ac.kr, kjkim@gist.ac.kr

GIST, South Korea



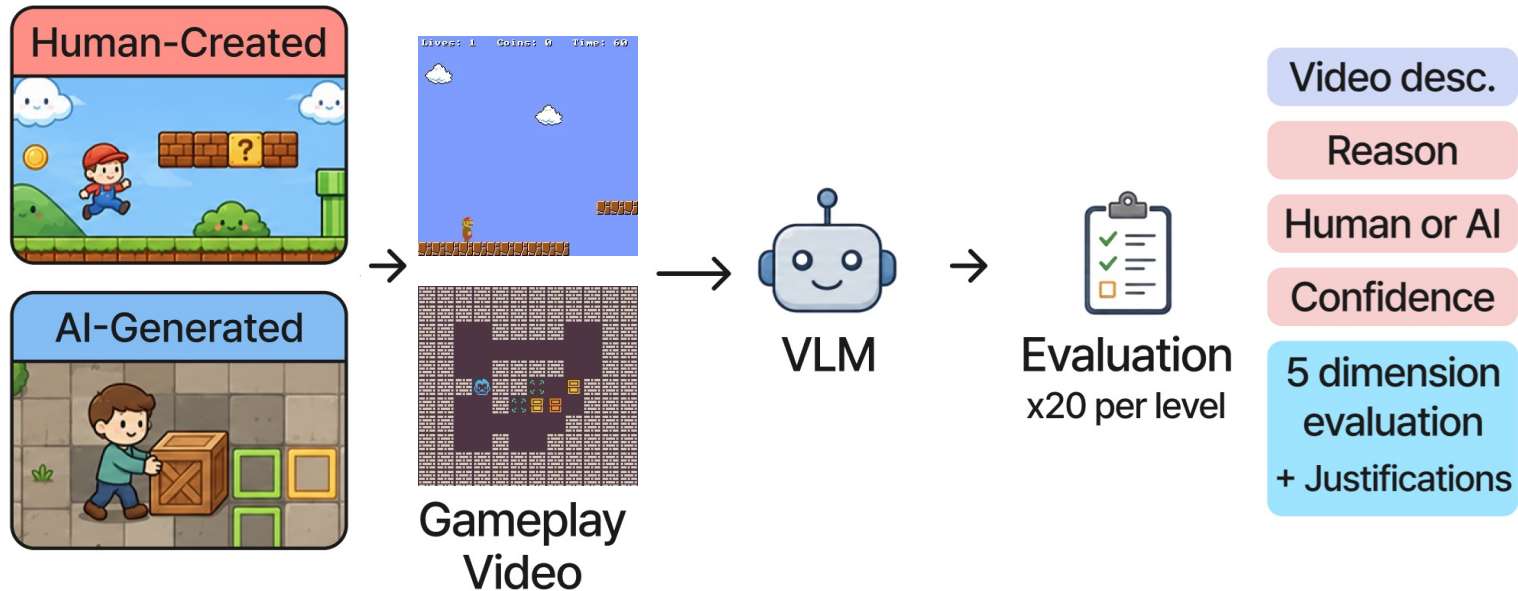
Rank 	Rank Spread 	Model 	Score 	95% CI ($\pm$)
1	1  7	 gemini-3-pro	1265	± 39
2	1  5	 gemini-2.5-pro	1258	± 23
3	1  14	 gpt-4.1-2025-04-14	1218	± 26
4	1  14	 gemini-2.5-flash	1210	± 26
5	2  15	 gpt-5-chat	1208	± 26
6	1  16	 chatgpt-4o-latest-20250326	1204	± 34
7	3  16	 gpt-4.1-mini-2025-04-14	1195	± 28
8	3  17	 o4-mini-2025-04-16	1190	± 26
9	3  17	 gemini-2.5-flash-lite-preview-06-17-thinking	1188	± 26
10	3  18	 o3-2025-04-16	1184	± 25

The Previous Study

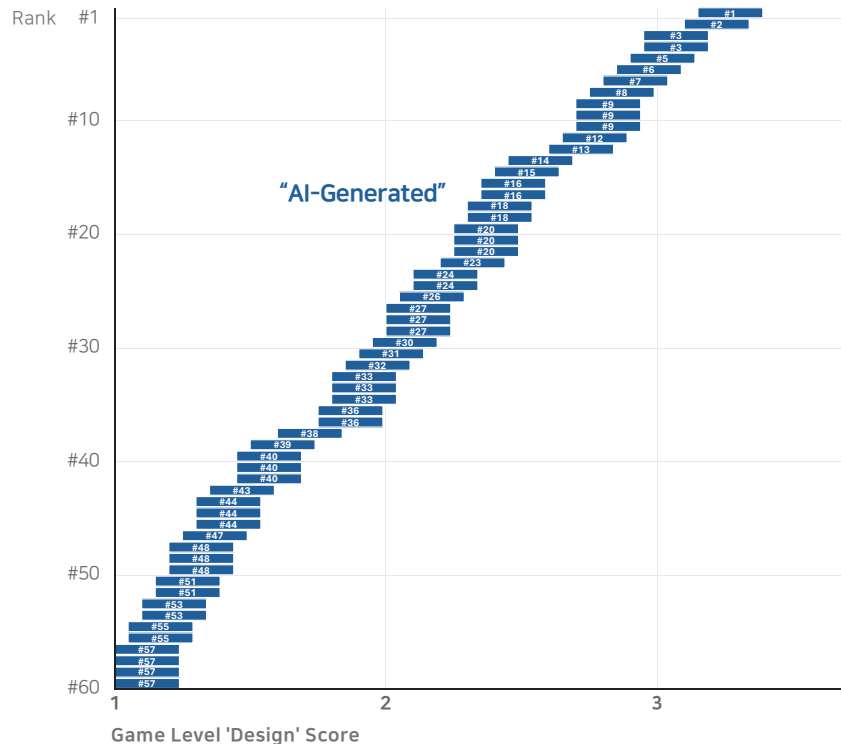


Fun, Challenging, Frustrating,
Surprising, Design

This Study (Experiment 1)



Results



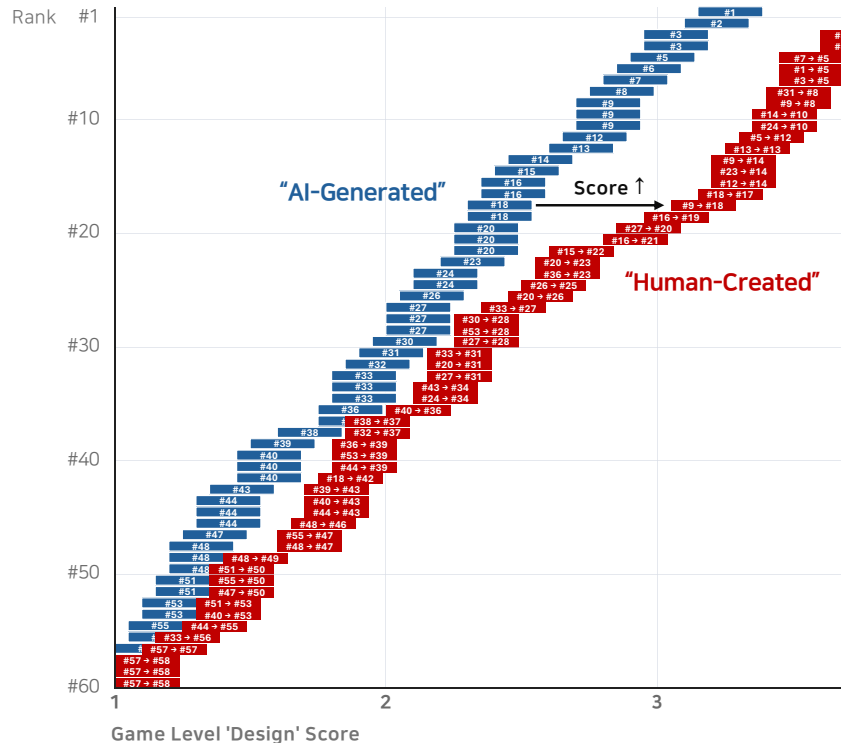
► Score Inflation

Stating "Human-Created" induced a systematic score increase across all metrics for identical levels.

► Consistent Relative Rankings

Although absolute scores shifted due to bias, the relative ordering of levels remained highly consistent across all evaluation metrics.

Results



► Score Inflation

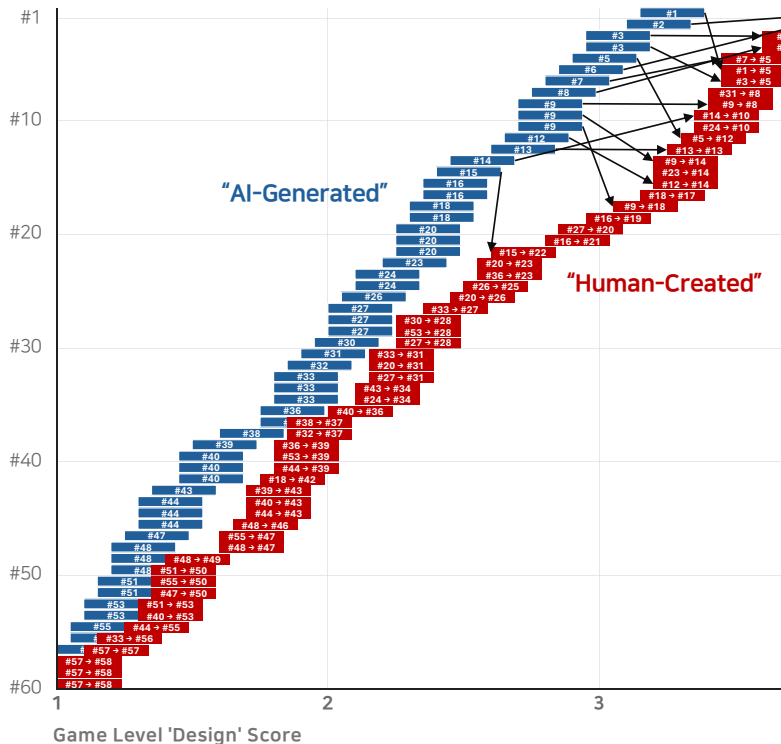
Stating "Human-Created" induced a systematic score increase across all metrics for identical levels.

► Consistent Relative Rankings

Although absolute scores shifted due to bias, the relative ordering of levels remained highly consistent across all evaluation metrics.

Results

Rank #1



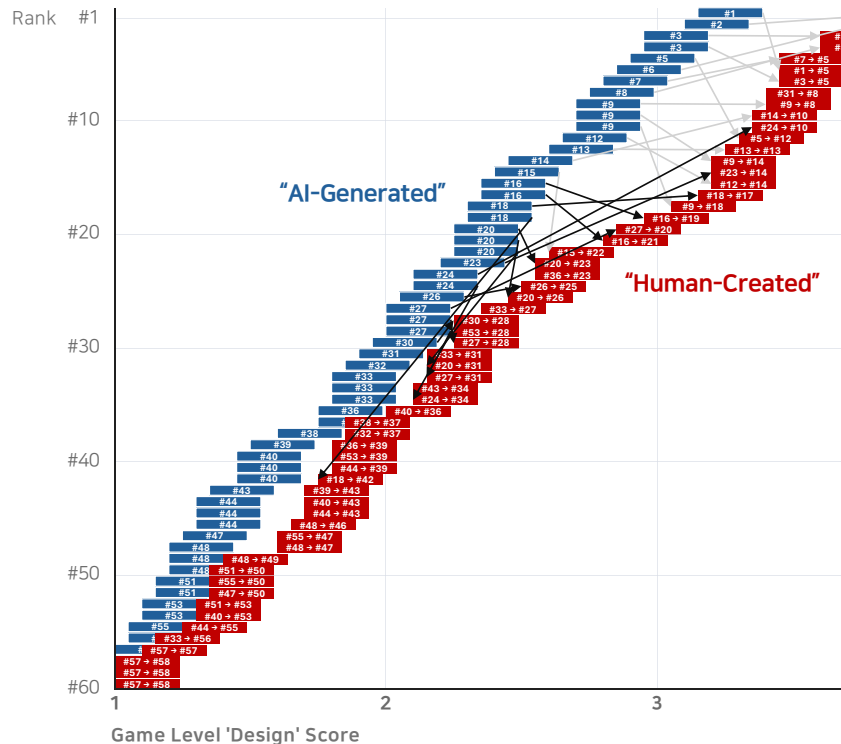
► Score Inflation

Stating "Human-Created" induced a systematic score increase across all metrics for identical levels.

► Consistent Relative Rankings

Although absolute scores shifted due to bias, the relative ordering of levels remained highly consistent across all evaluation metrics.

Results



► Score Inflation

Stating "Human-Created" induced a systematic score increase across all metrics for identical levels.

► Consistent Relative Rankings

Although absolute scores shifted due to bias, the relative ordering of levels remained highly consistent across all evaluation metrics.