AI Score Hallucination in Vision-Language Models

Measuring Self-Assessment Calibration for Iterative 3D Scene Positioning

TYLER VARACCHI
DREXEL UNIVERSITY

SIGGRAPH 2026
Los Angeles 19–23 JUL

MAIN FINDING

Across three vision-language models, self-reported confidence (~84/100) overstated real accuracy (47%) by 37.2 points and stayed statistically indistinguishable across models (ANOVA p = 0.87), despite accuracy ranging from 17% to 83%. I characterize this AI Score Hallucination.

AI confidence does not track accuracy

SCAN HERE
for a virtual poster tour with Tyler Varacchi

PROBLEM

VLM-driven pipelines use AI self-scores to decide when to stop iterating.

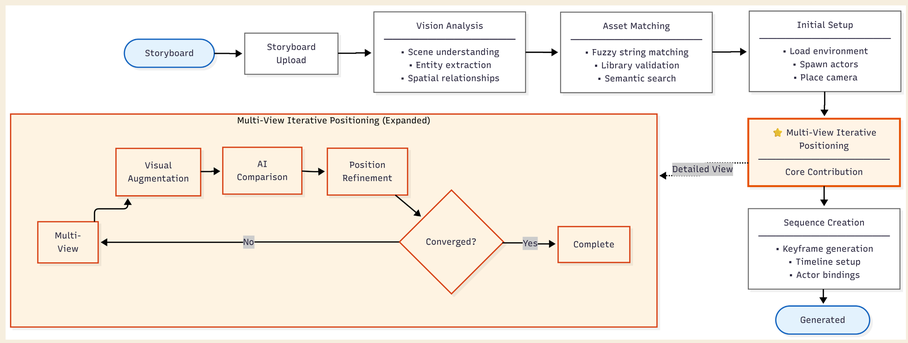
If those scores don't reflect real scene quality [2-4], loops fail invisibly: stopping too early or burning compute on completed scenes.

Silent failures pass AI review.

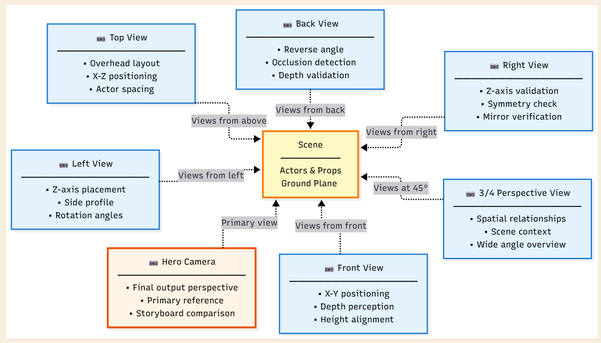
METHOD

PROCESS: Each iteration captures the 3D scene from seven viewpoints, sends them to the selected VLM, and applies the returned positioning adjustments.

CONTROL: all three VLMs received byte-identical prompts; only the VLM backend differed across all 36 scenarios; final scenes scored by one expert rater [5].



System architecture: storyboard panel to positioned 3D scene

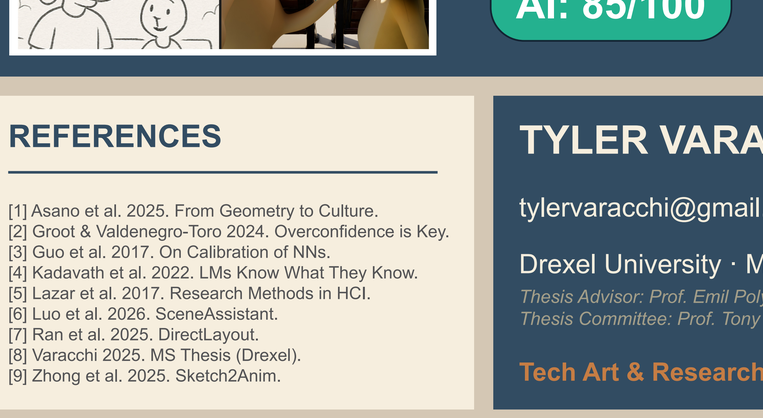


Seven viewpoints captured and scored per scene

MY APPROACH

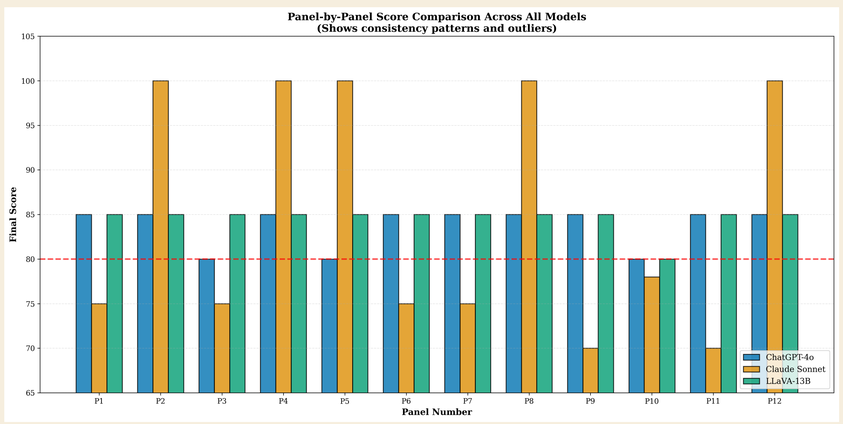
StoryboardTo3D plugin · C++ + embedded Python · Unreal Engine 5.6 [8]

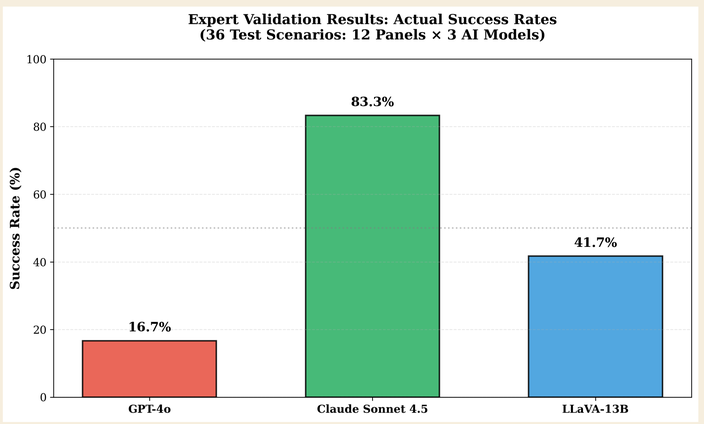
- Per panel: detect entities, match a project asset library, place the scene
- Each pass, the VLM re-scores its own scene 0 to 100 and adjusts placement
- Stop at self-score >80% strict, or a 30-iteration safety cap
- Mean iterations/scene: 7.8 LLaVA · 12.5 ChatGPT-4o · 22.2 Claude
- Extends VLM-as-judge layout and sketch-to-3D pipelines [1, 6, 7, 9]



The StoryboardTo3D plugin interface in Unreal Engine 5.6

RESULTS





PER-MODEL CALIBRATION ERROR

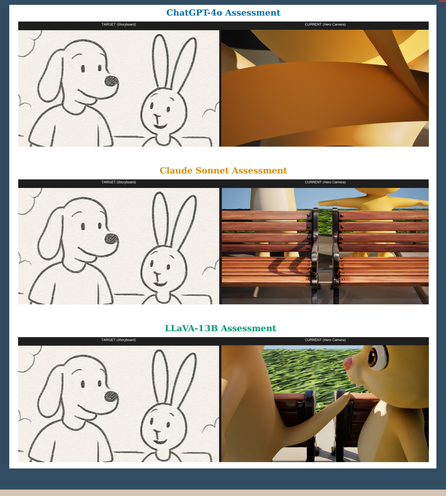
Claude Sonnet 4.5 ET	**+1.5 pp**	*(best · bimodal)*
LLaVA-13B	**+42.9 pp**	*(moderate · narrow SD)*
ChatGPT-4o	**+67.1 pp**	*(severe · narrow SD)*

- ChatGPT-4o and LLaVA-13B self-scored ~80 to 85 on every panel; only Claude varied
- Real accuracy splits 16.7% / 41.7% / 83.3%
- Calibration error: +1.5 to +67.1 pp
- Self-score is not a usable stop signal

PANEL 9 CASE STUDY

Same scene. Three VLMs.

ALL FAILED VALIDATION



ChatGPT-4o
AI: 85/100

Claude
AI: 70/100

LLaVA-13B
AI: 85/100

IMPLICATIONS

1. **PRODUCTION FIX:** Pair narrow-distribution VLMs with external validators (geometric checks · second-model judges · human review). Do not trust self-scores alone.
2. **STATISTICAL REALITY:** Self-scores are indistinguishable across models (ANOVA p = 0.87). The 37.2 pp accuracy gap is INVISIBLE without external validation.
3. **MODEL DIFFERENCE:** Only Claude is bimodal (SD 13.56), enabling thresholds. ChatGPT-4o (SD 2.26) & LLaVA-13B (SD 1.44) cluster 80 to 85 regardless of quality.
4. **LIMITATIONS:** single expert rater · cartoon corpus · static scenes · Oct-2025 model snapshots

REFERENCES

[1] Asano et al. 2025. From Geometry to Culture.
[2] Groot & Valdenegro-Toro 2024. Overconfidence is Key.
[3] Guo et al. 2017. On Calibration of NNs.
[4] Kadavath et al. 2022. LMs Know What They Know.
[5] Lazar et al. 2017. Research Methods in HCI.
[6] Luo et al. 2026. SceneAssistant.
[7] Ran et al. 2025. DirectLayout.
[8] Varacchi 2025. MS Thesis (Drexel).
[9] Zhong et al. 2025. Sketch2Anim.

TYLER VARACCHI

tylervaracchi@gmail.com

Drexel University · MS Digital Media (2025)
Thesis Advisor: Prof. Emil Polyak
Thesis Committee: Prof. Tony Rowe, Prof. Michael Wagner

Tech Art & Research Engineer

PORTFOLIO · CODE & DEMO · FULL THESIS · ACM PAPER