Measuring confidence gating for multimodal decision support under 46% visual-text disagreement: a threshold audit

ABSTRACT
We evaluated confidence gating for multimodal decision support under 46% visual-text disagreement. A deterministic paired simulation generated 64 cases and preserved a median slice. Mean decision consistency changed from 0.522 to 0.558; the paired difference was +0.036 (95% interval +0.034 to +0.038). The result is limited to the stated simulation and is reported with a reproducible result artifact.

Keywords multimodal decision support; confidence gating; visual-text disagreement; paired simulation; reproducibility

1. Introduction

Method comparisons in multimodal decision support need an adverse slice as well as an overall average. The experiment focuses on Seeing Is Not Deciding: Can Multimodal LLMs Act as Effective CEOs? arXiv preprint arXiv:2608.05864. Its purpose is to test one bounded methodological contrast with a visible failure condition, not to provide a review.

The primary question is whether confidence gating improves decision consistency while retaining the direction of change in the median slice. The operating setting is burst-load; the stress level is fixed at 46% before analysis.

2. Methods

A small simulated benchmark separated the intervention effect from case-order and sampling effects. We generated 64 cases with seed 50030. Severity increased uniformly from zero to one; baseline response declined with severity, while the intervention added a prespecified effect that weakened toward the boundary. The complete formula, parameters, case values, and summary are stored in `experiments/01-R05-030.json`.

Reference [1] is used in Methods, not only as background: its work on Seeing Is Not Deciding: Can Multimodal LLMs Act as Effective CEOs? arXiv preprint arXiv:2608.05864 defines the focal mechanism represented by the confidence gating intervention.

For the stress range, [2] supplies abstract-backed evidence on multimodal, agent through the study Intelligent Disassembly System for PCB Components Integrating Multimodal Large Language Model and Multi-Agent Framework. Its evidence ledger records the specific descriptors decommissioned, incorporated, localization, orchestrates, consumption, destructive, disassembly, electricity, identifying, improvement, bottleneck, converting, electrical, electronic, validating, connected, dexterity, equipment; we map those archived descriptors to the visual-text disagreement control. For the error slice, [3] supplies abstract-backed evidence on decision through the study Decision making in large-scale language testing: Intersections of policy, practice and research. Its evidence ledger records the specific descriptors intersections, professionals, administered, intricacies, formulated, illustrate, intentions, nationwide, particular, university, intricate, revisions, undergone, adopting, involved, mainland, overview, practice; we map those archived descriptors to the visual-text disagreement control. For the reporting rule, [4] supplies abstract-backed evidence on multimodal, decision, agent through the study Med-AgentX: Multimodal Large Language Model Agents with Explainable Reinforcement Learning for Trustworthy Biomedical Decision Support. Its evidence ledger records the specific descriptors explainability, accountable, brittleness, explanation, biomedical, crossmodal, persistent, rationales, remarkable, clinician, expertise, fidelity, powerful, theloop, agentx, aligns, black, fuses; we map those archived descriptors to the visual-text disagreement control. For the baseline, [5] supplies abstract-backed evidence on decision, agent through the study Optimization of Decision Making Capabilities of Intelligent Characters in Strategy Games Based on Large Language Model Agent. Its evidence ledger records the specific descriptors characters, collecting, beginning, dialogues, heuristic, imitating, pokellmon, dettmers, matchups, modified, official, realizes, showdown, battles, example, georgia, letting, llamma3; we map those archived descriptors to the visual-text disagreement control. For the measurement, [6] supplies abstract-backed evidence on multimodal, decision through the study Hybrid Cnn-gru Model for Real-time Multimodal Decision-making in Image and Text Analysis. Its evidence ledger records the specific descriptors synchronized, synchronous, monitoring, tensorflow, confirms, exploits, pytorch, matlab, python, keras, lower, late, shap, architectures, combination, industrial, sequential, optimal; we map those archived descriptors to the visual-text disagreement control. For the stress range, [7] supplies abstract-backed evidence on multimodal, decision, agent through the study Large Language Model-Enhanced Agent-Based Modeling for Intelligent Crowd Evacuation under Disaster Scenarios. Its evidence ledger records the specific descriptors probabilistic, assumptions, emergencies, populations, dependence, evacuation, community, dependent, histories, represent, specified, vehicular, allowing, exchange, indicate, movement, hazard, batch; we map those archived descriptors to the visual-text disagreement control. For the error slice, [8] supplies abstract-backed evidence on multimodal, decision through the

study A RAG-Enhanced Multimodal Large Language Model Framework for Culvert Inspection in Utah. Its evidence ledger records the specific descriptors prioritization, transportation, assessments, inspections, operational, proprietary, department, inspection, alongside, embedding, exemplars, onevision, thousands, agencies, convnext, labeling, reliance, selected; we map those archived descriptors to the visual-text disagreement control. For the reporting rule, [9] supplies abstract-backed evidence on decision through the study Artificial intelligence derived large language model in decision-making process in uveitis. Its evidence ledger records the specific descriptors misinterpreted, ophthalmology, inflammatory, consecutive, correctness, descriptive, incorrectly, intraocular, classified, inaccurate, invaluable, ophthalmic, parametric, references, sufficient, weaknesses, correctly, databases; we map those archived descriptors to the visual-text disagreement control. For the baseline, [10] supplies abstract-backed evidence on multimodal, decision through the study From Street-View Imagery to Auditable Spatial Evidence: A Multimodal Large Language Model Workflow for Urban-Renewal Screening. Its evidence ledger records the specific descriptors encroachment, association, directional, preliminary, enclosures, facilities, geographic, occurrence, priorities, stratified, translated, comprises, producing, crossing, develops, elements, exceeded, historic; we map those archived descriptors to the visual-text disagreement control.

3. Experimental Protocol

The primary endpoint was decision consistency. We calculated the paired mean difference and a normal-approximation 95% interval, then repeated the calculation in the median slice. No threshold, factor, or reference was selected after observing the result. The threshold audit used the same case order for both arms.

Data provenance for multimodal decision support is synthetic and explicit: the local generator uses the published seed, burst-load setting, and parameter table; it does not claim observed field measurements. This keeps the visual-text disagreement experiment inexpensive while preserving a rerunnable threshold audit.

4. Results

The baseline mean was 0.522; the intervention mean was 0.558. The paired change was +0.036, with a 95% interval from +0.034 to +0.038. In the prespecified adverse slice, the change was +0.032.

The machine-readable artifact `experiments/01-R05-030.json` contains all 64 paired records for multimodal decision support. Consequently, the +0.036 result can be recomputed directly under the burst-load setting rather than inferred from prose.

5. Discussion

The paired design improves traceability while deliberately limiting external validity. For multimodal decision support under burst-load, the sign of the decision consistency change remained positive in both the full set and the median slice. This supports a focused follow-up on visual-text disagreement using measured inputs and the same threshold audit endpoint.

For multimodal decision support, the references serve distinct roles: the locked target work defines confidence gating, while the abstract-backed supplements define visual-text disagreement, the median slice, and the threshold audit controls. None is included only to reach the ten-reference minimum.

6. Limitations and Conclusion

The multimodal decision support simulation is intentionally small and simplified. It omits acquisition bias, unmeasured confounding, hardware variability, and the deployment cost of confidence gating. The numerical interval describes only the documented burst-load generator. A real-data study should retain visual-text disagreement and the median slice before making any substantive performance claim.

We conclude that confidence gating is testable for multimodal decision support under visual-text disagreement; the present contribution is the reproducible protocol and result artifact, not a claim of real-world superiority.

References

1. Dai, Y., Peng, X., Wang, Y., Nakov, P., & Xie, Z. (2026). Seeing Is Not Deciding: Can Multimodal LLMs Act as Effective CEOs? arXiv preprint arXiv:2608.05864. arXiv. https://doi.org/10.48550/arXiv.2608.05864

2. Wang, L., Ouyang, L., Weng, H., Chen, X., Wang, A., & Zhang, K. (2026). Intelligent Disassembly System for PCB Components Integrating Multimodal Large Language Model and Multi-Agent Framework. Processes, 14(2), 227. https://doi.org/10.3390/pr14020227

3. Jin, Y. (2023). Decision making in large-scale language testing: Intersections of policy, practice and research. Studies in Language Assessment, 64-92. https://doi.org/10.58379/hkkx5020

4. HIRAGAR, P. R. (2025). Med-AgentX: Multimodal Large Language Model Agents with Explainable Reinforcement Learning for Trustworthy Biomedical Decision Support. https://doi.org/10.21203/rs.3.rs-7773686/v1

5. Shi, K. (2024). Optimization of Decision Making Capabilities of Intelligent Characters in Strategy Games Based on Large Language Model Agent. Journal of Big Data and Computing, 2(3), 136-140. https://doi.org/10.62517/jbdc.202401320

6. Mustafayeva, A., Israfilova, E., Baxshiyeva, G., & Aslanova, S. (2026). Hybrid Cnn-gru Model for Real-time Multimodal Decision-making in Image and Text Analysis. https://doi.org/10.21203/rs.3.rs-9257523/v1

7. Yang, S., Zhang, Y., & Gu, C. (2026). Large Language Model-Enhanced Agent-Based Modeling for Intelligent Crowd Evacuation under Disaster Scenarios. https://doi.org/10.5194/egusphere-egu26-3725

8. Mohammadi, P., Asgari, S., Rashidi, A., & Adibpour Hassankiadeh, S. (2026). A RAG-Enhanced Multimodal Large Language Model Framework for Culvert Inspection in Utah. https://doi.org/10.2139/ssrn.6429799

9. Schumacher, I., Bühler, V. M. M., Jaggi, D., & Roth, J. (2024). Artificial intelligence derived large language model in decision-making process in uveitis. International Journal of Retina and Vitreous, 10(1), 63. https://doi.org/10.1186/s40942-024-00581-1

10. Zhou, M., & Yang, J. (2026). From Street-View Imagery to Auditable Spatial Evidence: A Multimodal Large Language Model Workflow for Urban-Renewal Screening. https://doi.org/10.2139/ssrn.7225622