Stable Evidence, Unstable Decisions: An Empirical Analysis of Model Decision Stability in Vision–Language Models

Ali Khoramfar[1], Mohammad Javad Dousti[1], Alireza Mohamadian[2], Heshaam Faili[1]
[1]Department of Electrical and Computer Engineering, College of Engineering,
University of Tehran, Tehran, Iran
[2]Advanced Diagnostic and Interventional Radiology Research Center (ADIR),
Tehran University of Medical Sciences, Tehran, Iran
[1]{khoramfar, mjdousti, hfaili}@ut.ac.ir , [2]alirezamohamadian.md@gmail.com

Abstract

VLMs provide visual information alongside their predictions, but it remains unclear whether consistency in such information implies consistent decisions. We study this question in a controlled medical-imaging setting using brain MRI with pathology-confirmed labels and expert lesion annotations. For each human subject and modality, we construct configurations that retain the lesion content while varying surrounding context and scale and measure decision flips together with consistency in model-reported influential slices. Across four diverse VLMs (including proprietary, open-source, and domain-specific models), flip rates reach up to 75% across lesion-containing presentations, often despite high overlap in reported evidence. When lesion-related content is removed, proprietary models rarely produce a categorical diagnosis, with abstention rates ranging from 63% to 99%. These results reveal a mismatch between reported evidence and decisions, motivating evaluation beyond accuracy. Our evaluation dataset is publicly available on Hugging Face[1].

1 Introduction

Large language models (LLMs) have achieved strong performance across a wide range of natural language processing tasks, but their text-only formulation fundamentally limits their ability to reason about visual information (Yin et al., 2024). Recent advances have addressed this limitation through Vision–Language Models (VLMs), which integrate visual perception and language modeling within a unified framework (Wu et al., 2023). By enabling joint reasoning over images and text, these models extend language-centric architectures to visually grounded settings (Zhang et al., 2024). As a result, VLMs are increasingly used in scenarios where model outputs extend beyond a single prediction to incorporate information grounded in the visual input.

As VLMs are increasingly applied in settings where decisions depend on visual information, questions of reliability and consistency naturally arise (Hartsock and Rasool, 2024). This is particularly important in high-stakes domains like healthcare, where altering the visual framing of an image might influence not only the model's diagnostic prediction but also the visual evidence it reports. Yet, how VLM decision behavior responds when the visually grounded content is presented in different ways remains insufficiently characterized (Michalkiewicz et al., 2025).

In this work, we analyze decision stability in VLMs by comparing predictions across controlled inputs that vary in visual presentation, while retaining task-relevant visual content, measuring decision flips and their relation to model-reported visual evidence signals.

To study this question in a setting that allows explicit control over visual information, we instantiate our study on brain MRI data. In this domain, salient image regions can be localized and manipulated in a structured manner, enabling systematic variation of visual presentation while ensuring the presence of annotated lesion regions.

Specifically, we construct a dataset featuring expert-defined region annotations that mark image areas expected to inform the decision. These annotations allow us to construct controlled input configurations that emphasize or withhold particular regions, providing a concrete testbed for analyzing how changes in visual presentation affect model decisions.

In the medical domain, model evaluation is often summarized by aggregate metrics such as accuracy. While useful, these metrics can miss aspects of model behavior affecting trust and reliability, motivating evaluation perspectives that

[1]https://huggingface.co/datasets/universitytehran/MRI-GBM-MET

Findings of the Association for Computational Linguistics: ACL 2026, pages 26153–26166

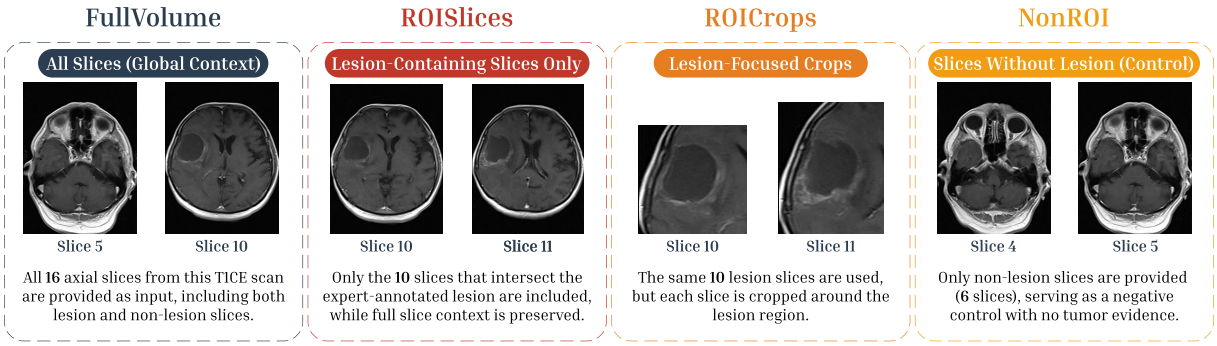

Figure 1: Illustration of input configurations for a single T1CE MRI subject. The patient has 16 axial slices in total, of which 10 intersect the expert-annotated lesion.

go beyond single-number scores and consider decision consistency (Alaa et al., 2025).

Motivated by these observations, we make the following contributions:

- **Proposed decision stability as a complementary evaluation dimension:** We show that aggregate accuracy alone is insufficient to characterize VLM behavior, as models can produce inconsistent decisions across alternative presentations of the decision-relevant visual content.
- **Suggested a controlled evaluation protocol:** We introduce a setup that systematically varies visual presentation while keeping the lesion content available, enabling direct analysis of decision consistency and abstention behavior.
- **Provided a curated brain MRI evaluation dataset:** We provide a curated brain MRI dataset with pathology-confirmed labels and expert lesion annotations, designed to support controlled research on presentation sensitivity and decision consistency.

The rest of this paper is organized as follows. Section 2 reviews related work, Section 3 outlines our dataset and experimental setup, and Section 4 and Section 5 present our results and conclusions, respectively.

2 Related Work

Recent large-scale evaluations reveal that standard accuracy metrics often mask fundamental fragility in VLMs. Rosenfeld et al. (2025) demonstrate that outputs fluctuate significantly under benign, meaning-preserving variations, arguing that stability is a critical predictor of reliability distinct from predictive performance. Similarly, Chou et al. (2025) empirically show that accuracy does not imply consistency; models frequently yield contradictory responses to semantically equivalent queries. This brittleness is exacerbated by cross-modal dependencies, where coordinated perturbations in both vision and language disrupt alignment more severely than unimodal noise (Babu et al., 2025).

A critical concern is the disconnect between model decisions and their attributed evidence. Liu et al. (2025) find that visual reasoning traces are often unfaithful, with predictions changing under textual interventions while ignoring visual manipulations. In the medical domain, Moll et al. (2025) report that answer accuracy and explanation quality are frequently decoupled, as models may rely on injected cues rather than true grounding. Furthermore, Han et al. (2025) identify a separation between "visual parsing" and "diagnostic reasoning," showing that models struggle to extract findings from images even when they possess strong text-based diagnostic logic.

While Michalkiewicz et al. (2025) analyze stability under viewpoint changes in foundation models, existing benchmarks largely focus on synthetic perturbations or reasoning faithfulness in isolation. There is limited analysis of decision stability when the clinical evidence is strictly preserved but presented in varying anatomical contexts. We address this gap by moving beyond noise-based robustness to evaluate consistency across controlled, lesion-containing input configurations. Furthermore, recent studies highlight the risk of strong textual biases dominating multimodal clinical AI models, often leading to the underutilization of visual evidence (Restrepo et al., 2026), underscoring the need for controlled evaluations.

Model	Mod.	Flip%			Ov@5		
		FullVolume vs. ROISlices	FullVolume vs. ROICrops	ROISlices vs. ROICrops	FullVolume vs. ROISlices	FullVolume vs. ROICrops	ROISlices vs. ROICrops
Gemini-2.5-Pro	T1CE	11.1	20.0	17.8	0.811	0.758	0.818
Gemini-2.5-Pro	T2	14.4	16.7	18.9	0.844	0.689	0.736
GPT-5.2	T1CE	7.8	18.9	13.3	0.876	0.833	0.860
GPT-5.2	T2	8.9	32.2	34.4	0.893	0.800	0.849
Qwen3-VL-32B	T1CE	60.0	60.0	2.2	0.420	0.427	0.931
Qwen3-VL-32B	T2	63.6	58.4	14.8	0.309	0.303	0.945
MedGemma-1.5-4B-IT	T1CE	35.8	61.0	53.2	0.719	0.551	0.618
MedGemma-1.5-4B-IT	T2	14.1	75.0	73.8	0.624	0.348	0.438
Random Baseline	T1CE	66.7	66.7	66.7	0.237	0.237	0.661
	T2	66.7	66.7	66.7	0.251	0.251	0.517

Table 1: Decision flips (Flip%) and agreement in model-reported influential slices (Ov@5) across presentation pairs. One FullVolume instance was excluded for MedGemma due to context length limits.

3 Methodology

3.1 Dataset and Task

Our study is conducted on a curated, retrospectively collected brain MRI dataset comprising 4,091 axial slice images acquired from 90 human subjects (all tumor-positive). The data were collected from four academic centers affiliated with Tehran University of Medical Sciences: Imam Khomeini Hospital Complex, Sina Hospital, Shariati Hospital, and Children's Medical Center. This cohort is distinct from public benchmarks, specifically assembled to prevent training–evaluation data contamination. Subjects are evenly balanced between glioblastoma (`GBM`, $n = 45$) and brain metastasis (`MET`, $n = 45$).

Ground-truth labels are pathology-confirmed. Bounding boxes and lesion-intersecting slices were derived from pixel-level consensus annotations by two board-certified neuroradiologists (see Appendix A for extended definitions and statistics).

We deliberately selected this differential diagnosis because it is a common yet genuinely challenging clinical task characterized by substantial intra-class heterogeneity (e.g., diverse primary origins for metastasis), providing a rigorous testbed for decision consistency under a non-trivial decision boundary.

Each subject includes two routinely acquired MRI modalities: T1-weighted contrast-enhanced MRI (T1CE) and T2-weighted MRI (T2). Expert-defined lesion masks enable explicit identification of lesion-intersecting slices, supporting controlled construction of lesion-containing visual inputs.

The task is sequence-level diagnosis: Given a set of slice images from a single modality (T1CE or T2) for one subject, the model predicts one of {`GBM`, `MET`, `UNSURE`} and reports a confidence score. Modalities are evaluated independently.

3.2 Input Configurations

For each subject and modality, we construct input configurations from the same underlying scan (Figure 1), varying visual presentation while retaining annotated lesion evidence. We refer to the expert-annotated *region of interest* (ROI) as the lesion region used to define ROI-intersecting slices and ROI-centered crops.

FullVolume. All axial slices from the sequence are provided in the order they appear in the scan. All provided slices were included in the model input, with no subsampling or image dropping applied.

ROISlices. Only slices intersecting the expert-defined lesion region are included, preserving full-slice context within those slices.

ROICrops. Only the slices intersecting the expert-defined lesion region are included (identical to the slice selection in ROISlices). However, for each of these slices, instead of full-slice context, only a localized crop centered on the lesion bounding box is provided, with a fixed margin and standardized resolution.

NonROI (negative control). Only slices without lesion annotations are provided, serving as a control condition where lesion evidence is absent.

3.3 Model Query Protocol

We evaluate four state-of-the-art VLMs as black-box systems: GPT-5.2 (OpenAI, 2025) and Gemini-2.5-Pro (Comanici et al., 2025) as proprietary generalists, Qwen3-VL-32B (Bai et al., 2025) as an open-source generalist, and MedGemma-1.5-4B-IT (Sellergren et al., 2026) as a medical-domain-specific model. For MedGemma, one FullVolume instance (160 slices) exceeded the maximum context length and was excluded from its evaluation. For each (subject, modality, configuration), the model receives the corresponding images and is prompted to output: (i) a predicted label in {GBM, MET, UNSURE}, (ii) a confidence score, and (iii) a ranked list of the top-K influential slice IDs, selected strictly from the slices provided in that configuration.

We treat the influential-slice ranking as a model-reported evidence signal and use it solely to assess self-consistency across input configurations. We make no claims about faithfulness or causality, and our analysis does not rely on attribution validity. All queries use a fixed prompt and low-variance decoding (temperature = 0), reducing sampling-induced variability.

Given the extreme flip rates observed in open-source models, our deeper consistency analyses focus primarily on the proprietary models.

3.4 Stability Metrics

We quantify stability at both the decision and reported-evidence levels. For each subject and modality, we compute the *decision flip rate* between two configurations as the fraction of subjects whose predicted label changes; we treat UNSURE as a valid output category, so flips include transitions between {GBM, MET} and UNSURE. We quantify consistency using Ov@K, the normalized overlap between the top-K slice sets reported under two configurations (K=5; for instance, if 4 out of 5 reported slices match, Ov@5 = 0.8).

We also report *Top-1 agreement*, the fraction of subjects where the most influential slice matches. Finally, to operationalize decision–evidence decoupling, we report flip rates conditioned on high Ov@5 agreement (e.g., Ov@5 $\geq$ 0.8). Since ROISlices and ROICrops share the lesion-intersecting slice pool, Ov@5 is best viewed as a within-pool consistency signal rather than a faithfulness guarantee.

Model	Mod.	UNSURE%	Conf≥0.8 & non-UNSURE%
Gemini-2.5-Pro	T1CE	63.3	34.4
Gemini-2.5-Pro	T2	84.4	8.9
GPT-5.2	T1CE	87.8	0.0
GPT-5.2	T2	98.9	0.0

Table 2: Negative control (NonROI): abstention behavior when lesion evidence is removed. UNSURE% is the fraction of subjects where the model abstains.

4 Results

We structure our evaluation around three core research questions: (i) how stable are model decisions across different lesion-containing presentation formats? (Section 4.1) (ii) does consistency in model-reported evidence align with consistency in final diagnostic decisions? (Section 4.2) and (iii) how do models behave when task-critical visual evidence is entirely removed? (Section 4.3)

4.1 Decision stability across lesion-containing presentations

Table 1 summarizes decision flip rates across configuration pairs. Our primary comparison is FullVolume vs. ROISlices: Both inputs contain lesion-intersecting evidence from the same scan, but differ in how the evidence is presented (all slices versus only lesion-intersecting slices). Even under this lesion-containing presentation change, we observe consistent decision variation across models and modalities. For proprietary models, flip rates range from 7.8% to 14.4%, indicating that diagnoses can be sensitive to presentation choices even when clinically relevant regions remain available.

This instability is substantially amplified in the open-source generalist model. Qwen3-VL exhibits flip rates up to 63.6%, indicating extreme sensitivity to contextual framing. In contrast, MedGemma—despite being substantially smaller—shows lower flip rates in the primary FullVolume vs. ROISlices comparison (14.1%–35.8%). Further inspection reveals that MedGemma's flips primarily consist of transitions between a categorical label and UNSURE, rather than direct GBM↔MET reversals. This suggests that medical fine-tuning may encourage more conservative behavior, shifting presentation-induced uncertainty toward abstention rather than contradictory diagnoses.

To contextualize these variations, the expected flip rate for uniform random guessing is 66.7%. While models generally fluctuate less than random chance, an analysis of flip severity (Appendix C) reveals that switches predominantly occur between correct and incorrect/abstain states, rather than between two incorrect labels. This directly impacts diagnostic reliability.

We additionally include ROICrops as a lesion-centered presentation that emphasizes the annotated region. Flips involving ROICrops can be even larger in some settings (e.g., reaching 34.4% for GPT-5.2 and up to 75.0% for MedGemma), showing that presentation changes that reframe the same underlying scan can materially affect the model's final decision. Notably, these effects are not confined to a single modality or model (Table 1), suggesting a general stability concern for VLM-style diagnostic use.

4.2 Reported evidence consistency and decision–evidence decoupling

Alongside decision flips, Table 1 reports agreement in model-reported influential slices via Ov@5. For the proprietary models, Ov@5 remains relatively high (0.689–0.893) across all pairs, indicating consistency in the model's reported evidence signal even when the predicted label changes. This co-occurrence of high Ov@5 with non-trivial flip rates reveals a consistent decoupling pattern in multimodal reasoning: models can consistently highlight and localize similar regions as evidence, yet still draw contradictory diagnostic conclusions when the surrounding visual context changes. Therefore, reported evidence stability alone does not guarantee decision stability.

This pattern is visible in the primary comparison as well: FullVolume vs. ROISlices yields measurable flip rates while maintaining strong evidence overlap (Ov@5 between 0.811 and 0.893 across the proprietary models). In other words, even when the evidence suggests substantial agreement about which slices matter, the diagnostic output can still change. This overlap is substantially higher than the random baseline expected by chance (e.g., ∼0.24 for FullVolume vs. ROISlices; see Appendix C).

This decoupling pattern starkly contrasts with the open-source models. For instance, Qwen3-VL exhibits high decision flip rates alongside substantially degraded evidence overlap, indicating a broader breakdown in multimodal alignment. MedGemma demonstrates moderate evidence overlap, further highlighting that while medical fine-tuning stabilizes reported evidence compared to generalist open-source models, it does not fully resolve presentation sensitivity.

4.3 Negative control: abstention when lesion evidence is removed

To validate that instability is not explained by indiscriminate guessing under arbitrary inputs, we evaluate the NonROI control condition where lesion-intersecting slices are excluded. Table 2 shows that the proprietary models predominantly abstain in this setting (UNSURE% = 63.3–98.9), consistent with the absence of lesion evidence. Confident non-abstaining predictions are rare under this control, indicating that high-confidence categorical decisions are not the default response when lesion evidence is removed. Additional uncertainty-aware analyses and confidence intervals are reported in Appendix C.

Qualitative inspection of the few reported influential slices in this condition reveals they consist of background-only images, serving as negative evidence rather than supporting a specific categorical prediction. Furthermore, flip rates for these models between lesion-present inputs and NonROI range from 80% to 93%, exceeding the random baseline (66.7%) and demonstrating a consistent shift toward abstention.

5 Conclusion

We examine decision stability in VLMs for brain MRI diagnosis under controlled, lesion-aware presentation changes across clinically plausible inputs. Across models and modalities, predictions vary across alternative inputs that preserve the lesion content, even when models highlight largely the same slices. This indicates that apparently stable evidence signals do not guarantee consistent decisions. In a negative-control condition that excludes lesion-intersecting slices, models predominantly abstain and rarely issue high-confidence non-abstaining diagnoses, consistent with the absence of lesion evidence. Taken together, these findings show that accuracy alone does not fully characterize model behavior, and that decision stability under presentation changes is especially critical in high-stakes clinical applications.

Limitations

Our study is a controlled stress test of decision stability under input presentation changes, and it has several limitations.

First, evaluating open-source and domain-specific VLMs on full MRI volumes is severely constrained by context length limits. For instance, evaluating MedGemma required dropping one large scan sequence. While proprietary models currently handle these long visual contexts, evaluating larger patient cohorts remains challenging for open-source architectures without employing subsampling heuristics, which would confound our controlled evaluation setup.

Second, our evaluation is necessarily black-box. The objective of this work is to characterize decision stability rather than explanation faithfulness. Although internal reasoning may vary across configurations, this cannot be directly examined with current VLM interfaces. We thus analyze the model-reported top-K influential slices as an output-level signal and focus on their consistency across input configurations.

Third, our presentation variants are designed as controlled, clinically plausible stress tests that preserve lesion-containing content. However, they do not exhaustively capture all sources of variability present in real-world imaging data (e.g., acquisition artifacts, site or scanner differences), which may interact with presentation choices in ways not examined here.

Ethical Considerations

The retrospective clinical data collection and the use of human-subject MRI scans were approved by the Institutional Review Board and Ethics Committee of Tehran University of Medical Sciences (Ethics Code: IR.TUMS.IKHC.REC.1401.251). No additional data were collected, and there was no interaction with patients.

This work is conducted strictly for research purposes and does not involve deployment or clinical decision-making. Prior to any computational analysis, rigorous de-identification was performed: all DICOM headers, metadata, and protected health information were strictly stripped, retaining only derived 2D image representations. Manual verification was conducted to ensure no personal identifiers remained, in full compliance with applicable privacy and data-protection regulations. The retrospective and de-identified nature of the dataset ensures that the study does not introduce new risks to participants.

The models evaluated in this work are not used to provide medical advice, diagnoses, or treatment recommendations. By explicitly demonstrating the fragility and decision instability of current VLMs under benign visual presentation changes, our findings actively serve to mitigate dual-use risks and over-reliance. This underscores that these models are not yet sufficiently reliable for clinical deployment, and our results are intended solely to inform research on model evaluation.

The dataset is publicly available for research use in de-identified form, in accordance with applicable institutional approvals and data-use requirements. Access is provided together with documentation describing the dataset's scope, intended research use, and known limitations. The dataset is not intended for clinical decision-making, diagnostic use, or commercial deployment, and appropriate terms of use are included to support responsible reuse and minimize the risk of misuse.

References

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. 2025. Medical Large Language Model Benchmarks Should Prioritize Construct Validity. *Preprint*, arXiv:2503.10694.

Ashwin Ramesh Babu, Sajad Mousavi, Vineet Gundecha, Sahand Ghorbanpour, Avisek Naug, Antonio Guillen, Ricardo Luna Gutierrez, and Soumyendu Sarkar. 2025. Coordinated Robustness Evaluation Framework for Vision-Language Models. *Preprint*, arXiv:2506.05429.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-VL Technical Report. *Preprint*, arXiv:2511.21631.

Shih-Han Chou, Shivam Chandhok, Jim Little, and Leonid Sigal. 2025. MM-r^3: On (in-)consistency of vision-language models (VLMs). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4762–4788, Vienna, Austria. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the Frontier

with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *Preprint*, arXiv:2507.06261.

Dae Hee Han, Eui Jin Hwang, Soon Ho Yoon, Hyungjin Kim, and Taehee Lee. 2025. Decoupling Visual Parsing and Diagnostic Reasoning for Vision–Language Models (GPT-4o and GPT-5): Analysis Using Thoracic Imaging Quiz Cases. *American Journal of Roentgenology*.

Iryna Hartsock and Ghulam Rasool. 2024. Vision-language models for medical report generation and visual question answering: a review. *Frontiers in Artificial Intelligence*, Volume 7 - 2024.

Zujing Liu, Junwen Pan, Qi She, Yuan Gao, and Guisong Xia. 2025. On the Faithfulness of Visual Thinking: Measurement and Enhancement. *Preprint*, arXiv:2510.23482.

Mateusz Michalkiewicz, Sheena Bai, Mahsa Baktashmotlagh, Varun Jampani, and Guha Balakrishnan. 2025. Not all Views are Created Equal: Analyzing Viewpoint Instabilities in Vision Foundation Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9113–9123.

Johannes Moll, Markus Graf, Tristan Lemke, Nicolas Lenhart, Daniel Truhn, Jean-Benoit Delbrouck, Jiazhen Pan, Daniel Rueckert, Lisa C. Adams, and Keno K. Bressem. 2025. Evaluating Reasoning Faithfulness in Medical Vision-Language Models using Multimodal Perturbations. *Preprint*, arXiv:2510.11196.

OpenAI. 2025. GPT-5 system card. Technical report, OpenAI. Released August 13, 2025.

David Restrepo, Ira Ktena, Maria Vakalopoulou, Stergios Christodoulidis, and Enzo Ferrante. 2026. On the Risk of Misleading Reports: Diagnosing Textual Biases in Multimodal Clinical AI. In *AI for Clinical Applications*, pages 320–330, Cham. Springer Nature Switzerland.

Amir Rosenfeld, Neta Glazer, and Ethan Fetaya. 2025. Questioning the Stability of Visual Question Answering. *Preprint*, arXiv:2511.11206.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, Liron Yatziv, Tiffany Chen, Bram Sterling, Kenneth Philbrick, Richa Tiwari, Yun Liu, Madhuram Jajoo, Chandrashekar Sankarapu, Swapnil Vispute, and 23 others. 2026. MedGemma 1.5 Technical Report. *Preprint*, arXiv:2604.05081.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023. Multimodal Large Language Models: A Survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-Language Models for Vision Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644.

A Extended Dataset and Clinical Details

This work is based on a curated brain MRI dataset constructed to support controlled analysis of diagnostic behavior under different visual presentations. The dataset consists of multi-slice axial MRI scans acquired as part of routine clinical imaging, with subject-level pathology-confirmed labels and expert-provided lesion annotations. The data are organized at the subject level, with each subject contributing multiple 2D axial slices per imaging modality.

A.1 Clinical Definitions

To provide context for readers unfamiliar with neuro-oncology, we briefly define the core clinical concepts used in this study:

MRI context. Magnetic Resonance Imaging (MRI) is the standard non-invasive imaging modality for brain tumor assessment due to its high soft-tissue contrast and flexibility in probing different tissue properties. In clinical neuro-oncology, diagnosis and differential characterization of tumors typically rely on a small set of conventional MRI sequences rather than specialized or advanced acquisitions. Accordingly, this dataset focuses on two routinely acquired modalities that provide complementary diagnostic information.

Lesion. In brain MRI, a lesion refers to a region of abnormal signal intensity that differs from normal brain parenchyma and corresponds to underlying pathology. In neuro-oncology, this typically includes the tumor mass and associated changes such as necrosis or peritumoral edema. In our study, lesion regions are defined using expert radiologist annotations and serve as the task-relevant imaging findings used to construct controlled input configurations.

T1-weighted contrast-enhanced MRI (T1CE). T1-weighted contrast-enhanced MRI is obtained after administration of a gadolinium-based contrast agent and highlights regions of blood–brain barrier disruption. In brain tumor imaging, T1CE is commonly used to visualize enhancing tumor components and internal structural heterogeneity. In the dataset, T1CE scans are provided as ordered axial slice sequences, accompanied by expert lesion annotations that identify slices intersecting the enhancing tumor region.

Quantity	Median (IQR)	Range (min–max)
T1CE slices / subject	20 (19–24)	15–160
T2 slices / subject	20 (18–22)	15–62
ROI slices / subject (T1CE)	8 (6–11)	3–34
ROI slices / subject (T2)	11 (8.25–13)	4–23
ROI slice fraction (T1CE)	0.39 (0.27–0.53)	0.15–0.85
ROI slice fraction (T2)	0.55 (0.40–0.67)	0.20–0.89

Table 3: Dataset summary statistics computed from the derived slice-level representations used as model inputs. ROI slice denotes a slice for which a corresponding ROI crop exists, reflecting the final ROI definition used to construct ROISlices/ROICrops and the NonROI control.

T2-weighted MRI (T2). T2-weighted MRI is sensitive to differences in tissue water content and is routinely used to characterize non-enhancing tumor components and surrounding tissue effects. In particular, T2 images often capture peritumoral signal changes that extend beyond the contrast-enhancing core. The dataset includes full axial T2 slice sequences for each subject, with expert annotations indicating lesion-associated regions in this modality.

Glioblastoma (GBM) vs. Brain Metastasis (MET). Glioblastoma is the most common primary malignant brain tumor in adults, arising from glial cells and characterized by infiltrative growth. Brain metastases are secondary intracranial tumors originating from systemic malignancies (e.g., lung or breast cancer). Although pathologically distinct, their MRI appearance—particularly in solitary enhancing lesions—can be highly similar, making radiologic differentiation clinically challenging.

Label Space Design and Abstention. Because the dataset consists entirely of pathology-confirmed tumor-positive subjects, the clinical objective is strictly differential diagnosis (GBM vs. MET) rather than tumor detection. Consequently, the label space is restricted to GBM, MET, UNSURE. When visual evidence of the lesion is withheld (as in the NonROI condition), the input does not magically convert into a "healthy" scan; rather, it becomes a scan with insufficient evidence for differentiation. In such clinical scenarios, diagnostic abstention (UNSURE) is the safest and most appropriate response, a behavior that medically fine-tuned models like MedGemma successfully replicate.

A.2 Annotation Process

Lesion annotations were performed in a rigorous two-stage process. Two board-certified neuroradiologists independently delineated tumor regions for each subject and modality at the pixel level on axial slices to generate binary masks. Discrepancies were subsequently reviewed in a joint session and resolved through consensus discussion, resulting in a finalized consensus mask per case. All experiments and ROI extractions in this study are based exclusively on these consensus masks.

Subject-level diagnostic labels (`GBM` vs. `MET`) were determined strictly based on histopathological confirmation following surgical biopsy or resection, as recorded in the clinical data. Therefore, the ground-truth labels are entirely independent of the radiologist-drawn imaging masks.

In this work, the expert lesion masks are used solely to identify lesion-intersecting slices and to construct ROI-based input configurations. For the *ROICrops* configuration, crops are generated from these slices using a standardized bounding-box procedure with fixed padding and a minimum size constraint. This ensures consistent perilesional context across all cases while preventing arbitrary framing artifacts.

A.3 Dataset statistics

The dataset comprises 90 human subjects in total. All subjects are tumor-positive, with a strictly balanced ground-truth label distribution (45 `GBM`, 45 `MET`). Across all subjects, the dataset contains 2,199 T1CE slice images and 1,892 T2 slice images. Table 3 reports sequence lengths, ROI coverage, and lesion area statistics derived from the expert consensus masks. ROI slices are defined by the presence of an ROI crop, matching the final ROI used for ROISlices/ROICrops construction and the NonROI negative control.

B Experimental Setup and Reproducibility

B.1 System Prompt and API Details

To support reproducibility, Table 4 specifies the API snapshots and knowledge cut-off dates for the proprietary models evaluated in this study. All queries were executed with a fixed decoding temperature of 0 to minimize sampling-induced variability.

Model	Snapshot	Knowledge Cut-off
Gemini-2.5-Pro	Jun 2025	Jan 2025
GPT-5.2	Dec 2025	Aug 2025

Table 4: Evaluated proprietary API models, snapshot versions, and knowledge cut-off dates.

The system prompt provided to models is as follows:

Evaluation Prompt

You are a neuroradiology expert performing MRI-based differential diagnosis.
Input:
Axial MRI slices from a single patient, acquired from ONE modality only:
- T1-weighted contrast-enhanced (T1CE), or
- T2-weighted (T2)
Task:
Your goal is to differentiate between:
- Glioblastoma (GBM): a primary, infiltrative high-grade brain tumor
- Brain metastasis (MET): a secondary lesion, typically more circumscribed
Procedure:
1) Diagnosis
Classify the case as GBM, MET, or UNSURE based solely on the provided slices.
2) Decision confidence
Provide a confidence score for your decision as a float between 0.0 and 1.0.
3) Evidence identification
Select the TOP decision-critical slices (TOP_K will be specified, but you may receive fewer slices).
Return their slice_id values in ranked order.
Constraints:
- Output MUST be valid JSON and nothing else.
- Decision confidence must be a float in [0.0, 1.0].
- Use only the provided slice_id values exactly.
- Do not assume access to other MRI sequences, annotations, or clinical history.

B.2 Image Input Protocol and Context Limits

To ensure a rigorous and controlled multimodal evaluation, we strictly adhered to the following image input protocol:

- **Independent Processing:** Each axial slice was passed to the model as an individual, distinct image within a single multi-image prompt. We explicitly avoided image compositing techniques (e.g., mosaics or grid tiling), as these can distort native resolutions and introduce spatial artifacts.
- **Order Preservation:** Slices were provided in their exact anatomical sequence as acquired during the MRI scan, preserving the natural volumetric flow.

Model	Mod.	Config	Acc (95% CI)	SelAcc (95% CI)	Coverage (95% CI)
Gemini-2.5-Pro	T1CE	FullVolume	0.789 [0.700, 0.867]	0.789 [0.700, 0.867]	1.000 [1.000, 1.000]
Gemini-2.5-Pro	T1CE	ROICrops	0.767 [0.678, 0.856]	0.784 [0.697, 0.865]	0.978 [0.944, 1.000]
Gemini-2.5-Pro	T1CE	ROISlices	0.811 [0.733, 0.889]	0.811 [0.722, 0.889]	1.000 [1.000, 1.000]
Gemini-2.5-Pro	T2	FullVolume	0.678 [0.578, 0.778]	0.685 [0.589, 0.784]	0.989 [0.967, 1.000]
Gemini-2.5-Pro	T2	ROICrops	0.622 [0.522, 0.722]	0.636 [0.529, 0.730]	0.978 [0.944, 1.000]
Gemini-2.5-Pro	T2	ROISlices	0.667 [0.567, 0.767]	0.690 [0.595, 0.784]	0.967 [0.922, 1.000]
GPT-5.2	T1CE	FullVolume	0.611 [0.511, 0.711]	0.632 [0.535, 0.729]	0.967 [0.922, 1.000]
GPT-5.2	T1CE	ROICrops	0.689 [0.589, 0.778]	0.705 [0.602, 0.795]	0.978 [0.944, 1.000]
GPT-5.2	T1CE	ROISlices	0.667 [0.567, 0.767]	0.682 [0.580, 0.775]	0.978 [0.944, 1.000]
GPT-5.2	T2	FullVolume	0.578 [0.478, 0.678]	0.627 [0.519, 0.729]	0.922 [0.856, 0.978]
GPT-5.2	T2	ROICrops	0.533 [0.422, 0.644]	0.571 [0.466, 0.675]	0.933 [0.878, 0.978]
GPT-5.2	T2	ROISlices	0.600 [0.500, 0.700]	0.643 [0.537, 0.744]	0.933 [0.878, 0.978]

Table 5: Performance across configurations with 95% bootstrap confidence intervals (subject-level). Coverage denotes the fraction of non-UNSURE outputs.

- **Context Limits:** Our evaluated proprietary models (GPT-5.2 and Gemini-2.5-Pro) successfully processed up to 160 independent images in a single API call. However, open-source and domain-specific models like MedGemma-1.5-4B-IT possess stricter context window limits. For MedGemma, one FullVolume sequence containing 160 slices exceeded the maximum allowable tokens and was subsequently excluded to avoid confounding the evaluation via heuristic subsampling.

B.3 Confidence Score Generation

As part of the model query protocol, VLMs were instructed to output a self-reported confidence score (a float between 0.0 and 1.0). Because proprietary models operate as black boxes, this score is explicitly generated as text by the model rather than derived from token-level softmax probabilities or logits. We therefore treat it as a model-declared estimate rather than a rigorously calibrated posterior probability.

C Expanded Results and Additional Analyses

This appendix provides extended empirical analyses to support the main findings presented in section 4. We first establish the models' baseline capabilities on this task, then rigorously analyze their behavior when visual evidence is removed, compare their outputs to theoretical random baselines, and finally provide an in-depth analysis of the decoupling between reported visual evidence and diagnostic decisions.

C.1 Aggregate Performance vs. Instance-Level Stability

Before analyzing decision instability across alternative presentations, it is crucial to establish whether the models are capable of performing the diagnostic task, or if they are systematically failing on specific input formats (e.g., struggling inherently with cropped images).

To verify this, we compute baseline aggregate performance metrics—accuracy, selective accuracy, and non-abstaining coverage—across the three lesion-containing configurations. As shown in Table 5, the models maintain uniformly high coverage (92.2%–100%) across these inputs, indicating that categorical diagnostic reasoning remains active regardless of the presentation format.

Moreover, aggregate accuracy is broadly comparable across presentation formats within each model and modality. For example, the T1CE accuracy of Gemini-2.5-Pro shifts minimally from 78.9% under FullVolume to 81.1% under ROISlices, with heavily overlapping confidence intervals.

This contrast exposes a critical diagnostic illusion: aggregate performance metrics can appear highly stable even while patient-level decisions fluctuate significantly. While the total volume of correct predictions remains stable, the specific subset of subjects receiving correct diagnoses changes based on the visual presentation, underscoring the necessity of evaluating instance-level stability (section 4.1) rather than relying solely on aggregate accuracy.

Model	Mod.	UNSURE% (95% CI)	Conf≥0.7 & non-UNSURE% (95% CI)	Conf≥0.8 & non-UNSURE% (95% CI)
Gemini-2.5-Pro	T1CE	63.3 [53.3, 73.3]	36.7 [26.7, 46.7]	34.4 [25.5, 44.4]
Gemini-2.5-Pro	T2	84.4 [76.7, 91.1]	15.6 [8.9, 23.3]	8.9 [3.3, 14.4]
GPT-5.2	T1CE	87.8 [81.1, 94.4]	1.1 [0.0, 3.3]	0.0 [0.0, 0.0]
GPT-5.2	T2	98.9 [96.7,100.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]

Table 6: Negative control (NonROI) with 95% bootstrap confidence intervals (subject-level).

Model	Mod.	Pair	Flip%	Ov@5
Gemini-2.5-Pro	T1CE	NonROI vs. FullVolume	80.0%	0.09
Gemini-2.5-Pro	T1CE	NonROI vs. ROICrops	81.1%	0.00
Gemini-2.5-Pro	T1CE	NonROI vs. ROISlices	83.3%	0.00
Gemini-2.5-Pro	T2	NonROI vs. FullVolume	91.1%	0.02
Gemini-2.5-Pro	T2	NonROI vs. ROICrops	90.0%	0.00
Gemini-2.5-Pro	T2	NonROI vs. ROISlices	88.9%	0.00
GPT-5.2	T1CE	NonROI vs. FullVolume	90.0%	0.04
GPT-5.2	T1CE	NonROI vs. ROICrops	88.9%	0.00
GPT-5.2	T1CE	NonROI vs. ROISlices	88.9%	0.00
GPT-5.2	T2	NonROI vs. FullVolume	92.2%	0.04
GPT-5.2	T2	NonROI vs. ROICrops	92.2%	0.00
GPT-5.2	T2	NonROI vs. ROISlices	93.3%	0.00
Random Baseline (vs. FullVolume)			**66.7%**	**0.25**
Random Baseline (vs. ROISlices/ROICrops)			**66.7%**	**0.00**

Table 7: Stability metrics for the NonROI control condition compared against lesion-containing presentations.

C.2 Negative Control: Absence of Visual Evidence

Having established baseline performance, we expand upon the findings in section 4.3 to investigate model behavior under the NonROI negative control condition, where diagnostic lesion evidence is entirely removed. If the instability observed in lesion-present configurations were merely the result of arbitrary textual biases or random sampling, models would continue to output categorical diagnoses here.

Instead, models exhibit a strong, systematic shift toward abstention. As detailed in Table 6, GPT-5.2 abstains (UNSURE) in 87.8% of T1CE and 98.9% of T2 cases, yielding 0.0% confident categorical predictions. Gemini-2.5-Pro similarly abstains in the vast majority of cases. Furthermore, we observe that the self-reported confidence scores plummet; the rate of confident predictions (≥ 0.8) drops to near zero, indicating that the models' declared confidence accurately reflects their uncertainty when task-critical visual evidence is absent.

To understand the models' evidence attribution under these abstaining conditions, we qualitatively inspected the corresponding influential slice rankings. When predicting UNSURE, the models selected background-only slices lacking any radiologically apparent lesion-related features (e.g., contrast enhancement or edema). These slices appear to serve as negative evidence, corroborating the absence of diagnostic features rather than supporting a hallucinated categorical prediction.

Finally, Table 7 details the stability metrics when comparing the NonROI control against all lesion-containing inputs. As expected, evidence overlap (Ov@5) collapses toward zero since the candidate slice pools are disjoint. Crucially, the flip rates transition to 80.0%–93.3%. This systematic divergence from the lesion-present flip rates confirms that the models possess the capacity to recognize missing evidence, and the flips observed in the main text are meaningful responses to how available information is structured.

C.3 Theoretical Random Baselines

To firmly contextualize the empirical instability reported in section 4.1 and demonstrate that it differs fundamentally from random behavior, we calculated theoretical random baselines for all evaluated metrics.

Configuration Context	Metric	Random Baseline	Actual Range
Lesion-containing presentations	Decision Flip Rate	66.7%	7.8%–34.4%
FullVolume vs. NonROI	Decision Flip Rate	66.7%	80.0%–93.3%
T1CE: FullVolume vs. ROISlices	Ov@5	∼0.24	0.758–0.876
T1CE: ROISlices vs. ROICrops	Ov@5	∼0.66	0.818–0.860
T2: FullVolume vs. ROISlices	Ov@5	∼0.25	0.689–0.893
T2: ROISlices vs. ROICrops	Ov@5	∼0.52	0.736–0.849

Table 8: Comparison of theoretical random baselines against the actual VLM empirical results across different modalities and presentation configurations.

Modality	Pair	Ov@5	Top-1
T1CE	FullVol vs. ROISlices/ROICrops	0.24	0.05
T1CE	ROISlices vs. ROICrops	0.66	0.13
T2	FullVol vs. ROISlices/ROICrops	0.25	0.05
T2	ROISlices vs. ROICrops	0.52	0.10

Table 9: Expected theoretical random baselines for evidence agreement metrics, calculated per-subject and averaged.

For a 3-class classification task (GBM, MET, UNSURE), a purely random model guessing uniformly has an expected accuracy of 33.3%. The probability that two independent random draws match is $1/3$, yielding an expected random **Flip Rate of 66.7%**.

For evidence agreement metrics (Ov@5 and Top-1 Agreement), the baseline depends on the number of slices available for each patient (M_i). We calculated a simplified expected overlap lower-bound ($5/M_i$ for Ov@5, and $1/M_i$ for Top-1, assuming uniform independent sampling) for each subject individually and averaged these across the dataset (Table 9). The baselines are identical for comparisons involving *FullVolume* because the random selection is drawn from the same full slice pool. The baseline is higher for the *ROISlices vs. ROICrops* comparison because the selection pool is restricted to the smaller set of lesion-intersecting slices.

Table 8 explicitly compares these theoretical baselines against the actual ranges observed across our evaluated VLMs. While the models show flip rates (7.8%–34.4%) that are substantially lower than random guessing (66.7%) in lesion-present configurations, they also exhibit evidence overlap (Ov@5) that vastly exceeds random chance. This confirms that the models are actively anchoring their reported evidence in the visual context, making the occurrence of decision flips highly significant.

Model	Mod.	Correct to Incorrect	Incorrect to Correct	Incorrect to Incorrect
Gemini-2.5-Pro	T1CE	4–11%	7–9%	∼0%
Gemini-2.5-Pro	T2	7–11%	6–7%	0–2%
GPT-5.2	T1CE	1–6%	7–13%	∼0%
GPT-5.2	T2	3–20%	6–13%	0–1%

Table 10: Directional flip severity ranges (min–max percentage across the three lesion-present presentation pairs). Incorrect outcomes include both categorical errors and the UNSURE category.

C.4 Directional Flip Severity and Correctness

While overall flip rates highlight presentation sensitivity, not all decision flips carry the same clinical consequence. A transition from a correct diagnosis to an incorrect one (or to abstention) is substantially more problematic than a model switching between two incorrect predictions.

To analyze this, we performed a correctness-aware directional analysis of decision flips. As shown in Table 10, across models and modalities, flips rarely occur between two incorrect predictions (∼0–2%). Instead, presentation changes predominantly shift cases between correct and incorrect/UNSURE outcomes. This demonstrates that the observed instability directly dictates whether a case is diagnosed correctly, further validating the necessity of stability metrics beyond aggregate accuracy.

C.5 Detailed Analysis of Model-Reported Evidence

Expanding upon section 4.2, this section provides a deeper analysis of the relationship between the models' reported visual evidence and their final diagnostic decisions.

Model	Mod.	Pair	Ov@5 (95% CI)	Top-1 Agree. (95% CI)
Gemini-2.5-Pro	T1CE	FullVolume vs. ROICrops	0.758 [0.711, 0.802]	0.489 [0.389, 0.589]
Gemini-2.5-Pro	T1CE	FullVolume vs. ROISlices	0.811 [0.769, 0.851]	0.778 [0.689, 0.867]
Gemini-2.5-Pro	T1CE	ROISlices vs. ROICrops	0.818 [0.778, 0.853]	0.533 [0.433, 0.633]
Gemini-2.5-Pro	T2	FullVolume vs. ROICrops	0.689 [0.644, 0.731]	0.489 [0.378, 0.589]
Gemini-2.5-Pro	T2	FullVolume vs. ROISlices	0.844 [0.802, 0.880]	0.700 [0.611, 0.789]
Gemini-2.5-Pro	T2	ROISlices vs. ROICrops	0.736 [0.693, 0.776]	0.478 [0.367, 0.589]
GPT-5.2	T1CE	FullVolume vs. ROICrops	0.833 [0.789, 0.876]	0.533 [0.433, 0.633]
GPT-5.2	T1CE	FullVolume vs. ROISlices	0.876 [0.838, 0.909]	0.711 [0.611, 0.800]
GPT-5.2	T1CE	ROISlices vs. ROICrops	0.860 [0.827, 0.889]	0.533 [0.422, 0.633]
GPT-5.2	T2	FullVolume vs. ROICrops	0.800 [0.744, 0.851]	0.511 [0.411, 0.611]
GPT-5.2	T2	FullVolume vs. ROISlices	0.893 [0.844, 0.933]	0.756 [0.667, 0.844]
GPT-5.2	T2	ROISlices vs. ROICrops	0.849 [0.809, 0.884]	0.567 [0.467, 0.667]

Table 11: Evidence stability across presentation pairs (Ov@5 and top-1 agreement) with 95% bootstrap confidence intervals (subject-level).

Model	Mod.	Pair	Flip(all)	Flip | Ov@5 $\geq t$			n eligible		
				t=0.6	t=0.8	t=1.0	t=0.6	t=0.8	t=1.0
Gemini-2.5-Pro	T1CE	FullVolume vs. ROICrops	0.200	0.175	0.169	0.192	80	59	26
Gemini-2.5-Pro	T1CE	FullVolume vs. ROISlices	0.111	0.110	0.088	0.111	82	68	36
Gemini-2.5-Pro	T1CE	ROISlices vs. ROICrops	0.178	0.179	0.176	0.188	84	74	32
Gemini-2.5-Pro	T2	FullVolume vs. ROICrops	0.167	0.143	0.156	0.077	77	45	13
Gemini-2.5-Pro	T2	FullVolume vs. ROISlices	0.144	0.128	0.100	0.105	86	80	38
Gemini-2.5-Pro	T2	ROISlices vs. ROICrops	0.189	0.192	0.185	0.200	78	54	20
GPT-5.2	T1CE	FullVolume vs. ROICrops	0.189	0.179	0.162	0.171	84	74	41
GPT-5.2	T1CE	FullVolume vs. ROISlices	0.078	0.069	0.077	0.078	87	78	51
GPT-5.2	T1CE	ROISlices vs. ROICrops	0.133	0.126	0.114	0.122	87	79	41
GPT-5.2	T2	FullVolume vs. ROICrops	0.322	0.317	0.333	0.350	82	69	40
GPT-5.2	T2	FullVolume vs. ROISlices	0.089	0.081	0.082	0.051	86	85	59
GPT-5.2	T2	ROISlices vs. ROICrops	0.344	0.345	0.346	0.325	87	78	40

Table 12: Flip rates conditioned on evidence agreement thresholds (Ov@5). We report conditional flip rates and the number of eligible subjects at each threshold.

C.5.1 Consistency of Reported Visual Evidence

Given that instance-level decisions fluctuate despite stable aggregate accuracy, a natural hypothesis is that the model makes different decisions because it focuses on different slices across configurations.

To test this hypothesis, we provide confidence intervals for the consistency of the models' reported influential slices in Table 11. The mean Ov@5 ranges from 0.811 to 0.893 for the primary comparison (FullVolume vs. ROISlices), alongside high Top-1 Agreement. This confirms that, at the level of reported attribution, models successfully and consistently identify overlapping subsets of influential images, regardless of whether the surrounding context is modified.

C.5.2 Threshold Sensitivity and Evidence–Decision Decoupling

Building on this observed consistency, we investigate whether higher degrees of agreement in reported evidence lead to a corresponding increase in decision stability. We quantify this by measuring flip rates across increasingly strict overlap thresholds ($\geq 0.6, 0.8$, and 1.0), as detailed in Table 12.

The results reveal a plateauing relationship, suggesting a fundamental decoupling between the reported evidence trace and the final diagnosis. For instance, in Gemini-2.5-Pro (T2, FullVolume vs. ROISlices), the flip rate initially drops from 14.4% to 10.0% at Ov@5 ≥ 0.8, but then remains stable at 10.5% when requiring perfect (1.0) evidence agreement. This pattern indicates that while high

Model	Mod.	Pair	Flip (95% CI)	Flip | Ov@5≥0.8 (95% CI)	Flip | Ov@5=1.0 (95% CI)	n Ov≥0.8	n Ov=1.0
Gemini-2.5-Pro	T1CE	FullVolume vs. ROICrops	0.200 [0.122,0.278]	0.169 [0.082,0.276]	0.192 [0.043,0.364]	59	26
Gemini-2.5-Pro	T1CE	FullVolume vs. ROISlices	0.111 [0.056,0.178]	0.088 [0.029,0.161]	0.111 [0.024,0.229]	68	36
Gemini-2.5-Pro	T1CE	ROISlices vs. ROICrops	0.178 [0.111,0.256]	0.176 [0.091,0.270]	0.188 [0.061,0.323]	74	32
Gemini-2.5-Pro	T2	FullVolume vs. ROICrops	0.167 [0.089,0.244]	0.156 [0.059,0.262]	0.077 [0.000,0.250]	45	13
Gemini-2.5-Pro	T2	FullVolume vs. ROISlices	0.144 [0.078,0.222]	0.100 [0.038,0.169]	0.105 [0.025,0.216]	80	38
Gemini-2.5-Pro	T2	ROISlices vs. ROICrops	0.189 [0.111,0.278]	0.185 [0.091,0.302]	0.200 [0.048,0.400]	54	20
GPT-5.2	T1CE	FullVolume vs. ROICrops	0.189 [0.111,0.267]	0.162 [0.085,0.254]	0.171 [0.059,0.300]	74	41
GPT-5.2	T1CE	FullVolume vs. ROISlices	0.078 [0.033,0.133]	0.077 [0.025,0.141]	0.078 [0.018,0.163]	78	51
GPT-5.2	T1CE	ROISlices vs. ROICrops	0.133 [0.067,0.211]	0.114 [0.050,0.187]	0.122 [0.026,0.229]	79	41
GPT-5.2	T2	FullVolume vs. ROICrops	0.322 [0.233,0.422]	0.333 [0.227,0.441]	0.350 [0.205,0.500]	69	40
GPT-5.2	T2	FullVolume vs. ROISlices	0.089 [0.033,0.156]	0.082 [0.025,0.143]	0.051 [0.000,0.113]	85	59
GPT-5.2	T2	ROISlices vs. ROICrops	0.344 [0.256,0.444]	0.346 [0.246,0.453]	0.325 [0.189,0.472]	78	40

Table 13: Flip rates with 95% bootstrap confidence intervals (subject-level), including conditional flip rates under high evidence agreement (Ov@5).

evidence overlap may sometimes coincide with stability, it does not provide a reliable guarantee of consistent decision-making.

C.5.3 Decision Stability Under Perfect Evidence Agreement

To further expose the limits of this decoupling, we isolate cases where models demonstrate absolute consensus in their reported evidence. We compute conditional flip rates specifically for subjects where the top-5 influential slices are identical across both presentations (Table 13).

Strikingly, even perfect agreement in model-reported evidence does not consistently resolve decision instability. A prominent example is GPT-5.2 on the T2 modality (FullVolume vs. ROICrops), where the unconditional flip rate is 32.2%. Among the subset of 40 subjects where the exact same five slices are highlighted as primary evidence, the flip rate remains largely unchanged at 35.0%. These findings strongly suggest that presentation-induced decision flips are not merely a byproduct of shifts in the model's reported visual focus, but rather a deeper inconsistency in how visual information is mapped to final decisions.