# Prompts Without Evidence: How Neuroimaging Mentions Shift Clinical Vision-Language Model Predictions

**Anonymous ACL submission**

## Abstract

Trustworthy clinical AI must ground its performance in genuine evidence rather than surface-level artifacts. We evaluate 12 open-weight vision-language models (VLMs) on two clinical neuroimaging cohorts for binary classification of affective disorders and cognitive decline. Both include structural MRI collected incidentally for unrelated studies, whose neuroimaging markers carry no reliable individual-level diagnostic signal. Nevertheless, smaller VLMs gain up to 0.58 in F1 when neuroimaging context is introduced, becoming competitive with models an order of magnitude larger. Confidence estimation indicates the calibration improvement on those smaller models is driven mainly by the prompt-level MRI reference, even without images. Our expert preliminary clinical case study finds faithfulness remains low in every condition examined, with models introducing unverified clinical details. Finally, preference alignment suppresses MRI-referencing behavior but collapses performance toward random baseline, leaving the underlying failure mode unresolved. These results caution against reading surface metric gains as evidence of genuine multimodal integration, with direct implications for clinical VLM deployment.

## 1 Introduction

The application of vision-language models (VLMs) to clinical decision-making has attracted growing interest, with recent work exploring their use for diagnostic classification from multimodal patient data (Moor et al., 2023; Singhal et al., 2023; Li et al., 2023). A natural expectation is that richer inputs should improve performance only when supplying *meaningful* diagnostic evidence. While spurious correlations and hallucination in VLMs have received growing attention (Zhong et al., 2024; Howard et al., 2025), less attention has gone to how models respond when a clinically *related* but diagnostically *uninformative* modality is introduced. The distinction matters: a model that changes its behavior based on an individual uninformative modality cannot be trusted to meaningfully integrate clinical evidence. We study this phenomenon on two clinical cohorts, including both structured textual data and neuroimaging data collected incidentally for each patient. Prior work established that the neuroimaging markers carry no reliable individual-level diagnostic signal from neuroimaging (Winter et al., 2022, 2024). Furthermore, we exclude trivially discriminative features in consultation with domain experts, so that models cannot rely on trivially discriminative shortcuts to complete the task (classification of affective disorders and cognitive decline).

Across 12 open-weight VLMs, we show that mentioning neuroimaging in the prompt shifts predictions, with the largest gains in small models whose text-only behavior is near-degenerate. On FOR2107 (Kircher et al., 2019), `Qwen2.5-VL-3B` (Xu et al., 2025) improves by up to 0.58 F1, which lifts it from a near-degenerate baseline to the range of models an order of magnitude larger (e.g., `Qwen2.5-VL-72B`). This gain is not explained by added diagnostic evidence: the majority of it is driven by the *textual mention* of MRI availability, not by any image. Where adding an image yields further gains, these reflect image *presence* rather than content, as they persist when the real MRI is replaced by an unrelated out-of-domain image. We explore this textually-driven component via a phrase-level probe and a false-modality ablation on smaller models, and connect it to the broader modality-collapse phenomenon (Parcalabescu and Frank, 2023; Sim et al., 2025) in a clinical setting, linking it to work on prompt sensitivity (Lu et al., 2024; Ismithdeen et al., 2025) and priming effects (Jones and Steinhardt, 2022; Yoshida et al., 2025). We run a preliminary case study with a domain expert, analyzing the reasoning trace produced by `Qwen2.5-VL-3B`, the model

showing the most dramatic boost. We find that the shift produces confident, evidence-styled justifications decoupled from actual diagnostic inputs. Finally, we show that the problem cannot be fixed via preference alignment: on `Qwen2.5-VL-3B`, preference alignment fails to remove the effect, suggesting the effect may be a structural property of how these models process domain-specific multimodal prompts. Our contributions are:

- An evaluation of 12 open-weight VLMs on FOR2107 (Kircher et al., 2019) and OASIS-3 (LaMontagne et al., 2019), two cohorts with no reliable individual-level diagnostic signal.
- Confidence estimation and ablation analyses that associate the majority of the calibration improvement with textual preamble framing on smaller models, with the trigger further characterized via a phrase-level probe and a false-modality ablation.
- A preliminary expert case study with a clinical psychologist showing that, although multimodal context improves reasoning-trace quality, faithfulness remains below acceptable thresholds in every condition examined.
- A preference-alignment intervention showing that removing MRI-referencing behavior pushes all conditions toward chance rather than selectively fixing the modality dependence.

We analyze this phenomenon at three levels: *predictive* (F1 across conditions, Section 4.1), *behavioral* (confidence shifts and reasoning traces, Sections 5-6), and *representational* (the textual trigger and its response to alignment, Sections 7-8). These are not separate findings but three views of the same effect: a large gain in predictive performance that cannot be explained by added diagnostic evidence. Surface evaluations therefore overstate the diagnostic ability of VLMs in clinical settings, a risk that can go unnoticed before deployment.

## 2 Related Work

We review related work on prompt sensitivity, the priming effect and modality collapse in vision-language models (VLMs).

### 2.1 Prompt Sensitivity in VLMs

While prompt sensitivity has been extensively studied in large language models (LLMs) (Ishibashi et al., 2023; Lu et al., 2024), it remains underexplored in VLMs. Among the few studies on this, Ismithdeen et al. (2025) demonstrate that VLMs suffer from severe prompt sensitivity, a vulnerability that leads to inconsistent classification performance and a concerning reliance on language priors.

In healthcare, sensitivity to instruction phrasing can severely impact reliability: prompt variations cause F1 fluctuations of up to 0.25 in LLMs (Ceballos-Arroyo et al., 2024) and accuracy variations of up to 6% in VLMs on health benchmarks (Ismithdeen et al., 2025).

### 2.2 Priming Effect in VLMs

The priming effect is a cognitive phenomenon in which prior exposure to a stimulus influences subsequent judgments or behaviors, often without conscious awareness (Meyer and Schvaneveldt, 1971). Rather than altering decision-making through explicit alternative descriptions, priming operates by activating related concepts, associations, or response tendencies that shape how later inputs are interpreted (Koo et al., 2024).

In VLMs, priming refers to how prompts, both textual and visual cues, shape model outputs, mirroring the priming concept from cognitive science. Yoshida et al. (2025) directly investigate this phenomenon in large-scale VLMs, demonstrating that model responses systematically shift in the direction intended by an accompanying image, suggesting that VLMs actively incorporate visual information into language processing rather than treating it as incidental. Recent research has further revealed both the power and limitations of priming in adapting VLMs to new tasks and domains (Jones and Steinhardt, 2022; Koo et al., 2024). As these models increasingly incorporate multimodal inputs, such priming effects can be further intensified by interactions between textual and visual signals (Gulati et al., 2025; Zhang et al., 2026). For instance, Zhang et al. (2026) show that in fact-checking settings, VLMs may favor previously introduced textual context over salient visual evidence, and related behaviors have been observed in general VQA tasks (Shu et al., 2025).

### 2.3 Modality Collapse in VLMs

Modality collapse refers to the phenomenon where a unimodal model achieves comparable accuracy to a multimodal model on a vision-language task, revealing that one modality is not meaningfully utilized (Javaloy et al., 2022; Parcalabescu and Frank, 2023). A recent survey by Sim et al. (2025) provides a systematic taxonomy of contributing fac-

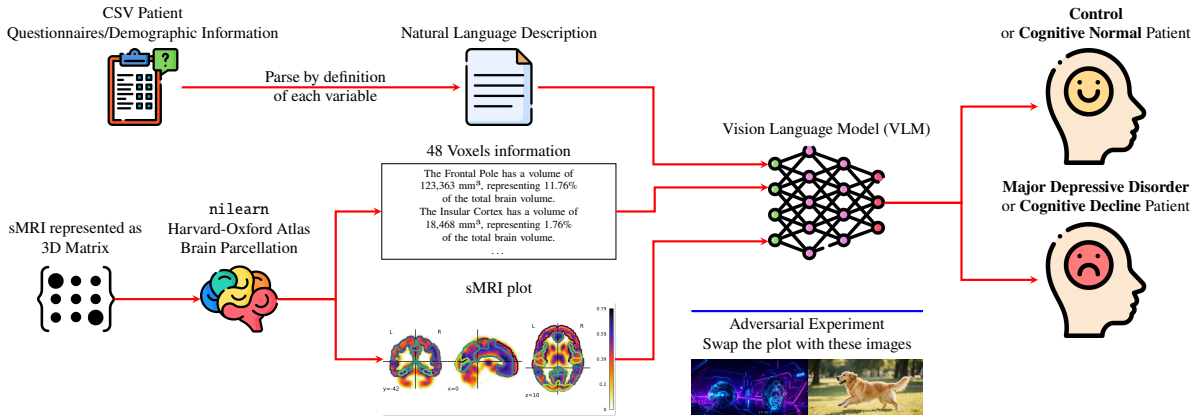



Figure 1: Overview of the proposed VLM pipeline.

tors, including dataset bias, model behavior, lack of fine-grained supervisory signal, and task setup, and reviews methods for quantifying modality contribution and cross-modal interaction. A consistent finding across this literature is that the text modality dominates, with visual input failing to influence predictions even when it is ostensibly required by the task (Zhu et al., 2022). In the clinical domain, this is referred to as "*shortcut learning*". Models often rely on imaging artifacts or dataset features rather than actual pathology (DeGrave et al., 2021; Geirhos et al., 2020).

Our work moves from detecting modality collapse to localizing its trigger. Prior modality collapse work establishes that text dominates. We instead identify *which part of the textual input* triggers the collapse in a clinical setting, and show that conventional suppression strategies cannot separate the trigger from the model's broader reasoning behavior.

## 3 Methodology

Our pipeline (Figure 1) integrates two primary modalities: **tabular clinical data** (CSV) and **structural brain MRI (sMRI)** scans. It has three sequential phases: (1) **Clinical Tabular Data Serialization:** We serialize patient-level CSV data into natural language by mapping each variable's value to its clinical description. This transformation has been shown to substantially improve comprehension and reasoning in language models (Hegselmann et al., 2023; Vu et al., 2025). (2) **sMRI Information Extraction:**[1] Following previous work (Kapoor and Egger, 2025; Galteau et al., 2025a,b), we use `nilearn`[2] and the Harvard-Oxford probabilistic cortical atlas (Rushmore et al., 2022) to visualize sMRI volumes and extract regional anatomical measurements via brain parcellation. Per-region volumetric measurements are serialized into descriptive text following the same approach as the tabular data. For visualization, we render three orthogonal slices (sagittal, coronal, axial) annotated with MNI coordinates, hemisphere labels, and voxel contour overlays. (3) **Multimodal Prediction:** The serialized text, parcellation descriptions, and MRI visualizations are aggregated into a structured chat template and forwarded through a VLM for binary classification. Input components can be included or removed at inference time, which defines the five conditions we evaluate (Section 4). The full prompt template is shown in Appendix A.

## 4 Experiment

**Datasets** We evaluate our pipeline on two clinical neuroimaging datasets:

**FOR2107** is a German multicenter cohort study on the neurobiology of affective disorders (Kircher et al., 2019). It includes patients with Major Depressive Disorder (MDD) and matched healthy controls, with deep phenotyping spanning structural MRI, clinical assessments, neuropsychological testing, and demographic information. Prior work on this cohort, despite a much larger sample, reports classification accuracies of only ≈56% with univariate neuroimaging markers (Winter et al., 2022) and a ceiling of 62% with multivariate ML over a 4M-model search across structural and functional MRI, DTI, a polygenic risk score, and environmen-

[1] T1-weighted sMRI scans in our cohorts are stored as 3D volumes in `.nii` or `.nii.gz` format, with common dimensions of 256 × 256 × 176-208 voxels.

[2] https://github.com/nilearn/nilearn

| Cond. | Input | csv | Prompt(MRI) | Parcel. | Image | Image Type |
|---|---|---|---|---|---|---|
| C1 | TEXT(CSV) | ✓ | | | | |
| C2 | TEXT(CSV) + PROMPT(MRI) | ✓ | ✓ | | | |
| C3 | TEXT(CSV) + PROMPT(MRI) + PLOT(MRI) | ✓ | ✓ | | ✓ | `nilearn` brain plot |
| C4 | TEXT(CSV, PARCEL) + PROMPT(MRI) + PLOT(MRI) | ✓ | ✓ | ✓ | ✓ | `nilearn` brain plot |
| C5 | TEXT(CSV, PARCEL) + PROMPT(MRI) + PLOT(SWAP) | ✓ | ✓ | ✓ | ✓ | OOD image (ablation) |

Table 1: Experimental conditions and their input components. Conditions are designed to isolate the contribution of each modality: C1 uses tabular text only, C2 adds the MRI preamble without imaging data, C3-C4 progressively add MRI content, and C5 substitutes a counterfactual image to control for image-driven effects.

tal risk factors (Winter et al., 2024). FOR2107 therefore represents a genuinely hard classification problem where performance gains warrant scrutiny. Our binary task distinguishes *Active MDD* from *Healthy Controls*.

**OASIS-3** is an open-access longitudinal dataset compiled from the Washington University Knight Alzheimer Disease Research Center (LaMontagne et al., 2019). It includes participants ranging from cognitively normal adults to individuals at various stages of cognitive decline, accompanied by multimodal MRI sessions and clinical assessments. To match the classification setup with FOR2107, we focus on target labels *Cognitive Decline* and *Cognitively Normal*.

Table 2 summarizes class distributions. Both datasets are governed by strict data-use agreements that prohibit redistribution and require formal application for access. To our knowledge, neither has appeared in any NLP or VLM publication, which makes training-data contamination unlikely. This provides a stronger control than is available with public benchmarks such as the 86 to 160 sample subsets used by Ceballos-Arroyo et al. (2024) or the 197 Health and Medicine examples in Ismithdeen et al. (2025). Both datasets contain many clinical variables drawn from multiple CSV files (full list in Appendix D and Appendix E). Some of these variables are easy classification shortcuts (e.g., suicidal thoughts). In consultation with domain experts in clinical psychology, we excluded such trivially discriminative features so that the tested VLMs cannot rely on trivially discriminative shortcuts to complete the task.

**Models** We test 12 popular open weight Vision-Language Models[3]: `InternVL3.5-4B` and `14B` (Wang et al., 2025), `GLM-4.1V-9B` (Hong et al., 2025), `GLM-4.6V-Flash` (Zeng et al., 2025), `LLaVA-OV-1.5-4B` (Li et al., 2025), `Ministral-3-3B` and `14B` (Liu et al., 2026), `Qwen2.5-VL-3B`, `32B` and `72B` (Xu et al., 2025), `Qwen3-VL-2B` and `32B` (Bai et al., 2025). For all models, we use the HuggingFace Transformers (Wolf et al., 2020) implementation[4], setting `do_sample=False` to ensure deterministic, reproducible outputs across runs. To contextualize model performance we also include a stratified random baseline on both datasets (derivation details in Appendix F).

| Dataset | Group / Status | # Samples |
|---|---|---|
| FOR2107 | Active MDD | 701 |
| | Control | 1,071 |
| OASIS-3 | Cognitive Decline | 487 |
| | Cognitive Normal | 849 |

Table 2: Class distribution of the clinical subsets used in this work.

**Naming Scheme** To evaluate the impact of clinically irrelevant input in our tasks, we evaluate five conditions progressively incorporating multimodal neuroimaging information into the pipeline, summarized in Table 1 (See Appendix A for the exact prompt):

- TEXT(CSV) consists of a textual representation of the patient metadata, usually collected in repeated sessions with the client by the clinical expert and used for diagnosis.
- PROMPT(MRI) adds a mention of MRI data (brain parcellation volume, visualization of brain regions) in the used prompt.
- TEXT(PARCEL) includes a textual representation of the brain parcellation data.
- PLOT(MRI) adds the MRI plot as image modality.
- PLOT(SWAP) replaces the brain plot with an out-of-distribution image (a dog photograph or styled sci-fi brain scan, see Appendix G) to ablate whether performance depends on image content

[3]We also conduct ablation experiments on the medical-specialized VLM `medgemma` (Sellergren et al., 2026), which also exhibits failure modes (Appendix H.2).

[4]For the detailed name, we refer to Table 8 in the Appendix.

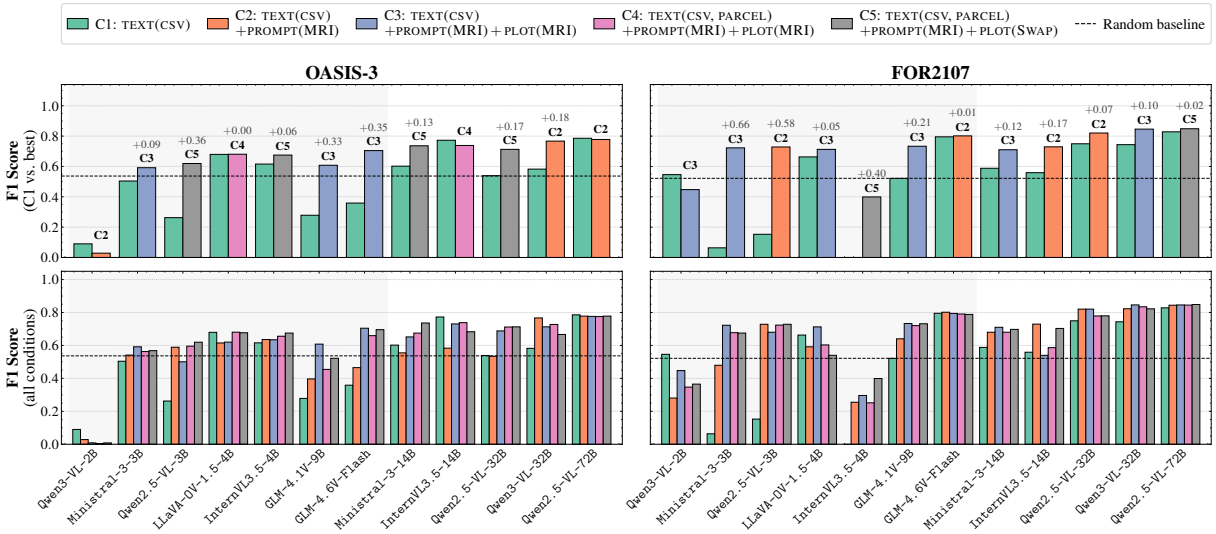



Figure 2: F1 scores on both datasets for the 12 tested VLMs. **Top row**: text-only baseline (C1, teal-green) versus the best augmented setup per model (best of C2-C5). **Bottom row**: full breakdown across all five input conditions. Models sorted from smallest to largest (parameter count).

| Cond. | Qwen2.5-VL-3B | | Ministral-3-3B | |
|---|---|---|---|---|
| | ECE↓ | Brier↓ | ECE↓ | Brier↓ |
| C1 | 0.376 | 0.370 | 0.361 | 0.354 |
| C2 | 0.164 | **0.178** | 0.199 | 0.219 |
| C4 | **0.155** | 0.184 | **0.183** | **0.207** |

Table 3: Per-condition calibration with 15 equal-width bins on predicted-class confidence ECE together with Brier. Lower ECE and Brier indicate better calibration. **Bold** indicates the best score per model.

or image presence alone.

### 4.1 Experiments Results

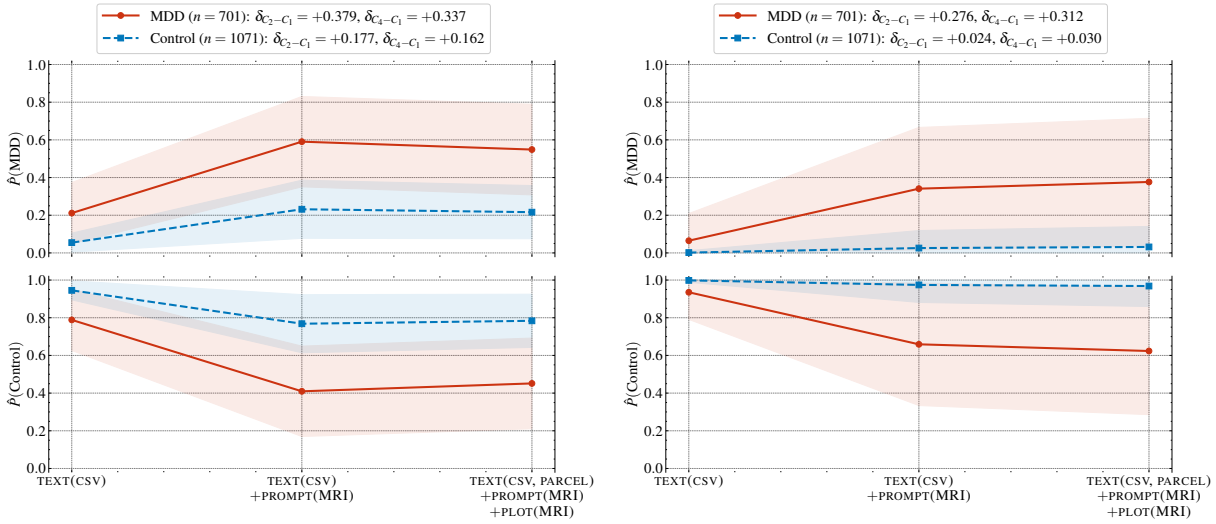
Figure 2 reports F1 across all models and conditions. Under the text-only baseline (C1), several smaller models fail to exceed the random baseline, most strikingly on FOR2107 where `Ministral-3-3B` (0.064) and `Qwen2.5-VL-3B` (0.153) fall far below it.

This pattern inverts once multimodal context enters: Figure 2 shows that smaller models gain the most, with `Qwen2.5-VL-3B` and `Ministral-3-3B` improving up to +0.58 and +0.66 F1 on FOR2107, while larger counterparts barely move. The swap-image condition (C5) shows that performance is largely preserved when the MRI plot is replaced by an unrelated image. From qualitative inspection, including neuroimaging context affects reasoning even when classification performance does not change: larger models produce MRI-referencing justifications under multimodal input despite stable F1.

## 5 Confidence Estimation

Since the smaller models exhibit high jumps in the F1 score, we further examine the probabilities the smaller models assign to each label. For each patient, we compute a normalized confidence score based on the model's joint probability of its full generated answer and the predicted label token, renormalized across $\{\text{MDD}, \text{Control}\}$ (formal definition in Appendix J). This formulation follows label-token probability extraction approaches (Zhao et al., 2021; Geng et al., 2024; Xiong et al., 2024). From these scores we measure *calibration* (Xiong et al., 2024): the degree to which the model's confidence agrees with its actual classification accuracy, captured by Expected Calibration Error (ECE) (Pakdaman Naeini et al., 2015; Guo et al., 2017) and the Brier score (Brier, 1950).

We report this analysis on `Qwen2.5-VL-3B` and `Ministral-3-3B`, the two smaller models with the largest C1→C2 gains on FOR2107 and thus the clearest cases for understanding where the gains come from. We also extend to `InternVL3.5-4B` and `GLM-4.1V-9B` in Appendix J.1, finding the same pattern. Table 3 reports ECE and Brier. For both models, conditions C2 and C4 reduce calibration error relative to the C1 baseline on both metrics. Predictions are thus not only more accurate but also better calibrated, indicating that the gains are not driven by uniform confidence inflation. We observe the bulk of this improvement




(a) `Qwen2.5-VL-3B` (b) `Ministral-3-3B`

Figure 3: FOR2107 per-cohort mean $\hat{P}(\text{MDD})$ (top) and $\hat{P}(\text{Control})$ (bottom) across the three input conditions. Both cohorts exhibit an upward shift in $\hat{P}(\text{MDD})$, with a mirror-symmetric downward shift in $\hat{P}(\text{Control})$.

under the MRI prompt alone (C2), i.e. before any imaging content is provided: the C1→C2 transition accounts for the majority of the total C1→C4 ECE reduction for both models, suggesting that the gains reflect contextual priming rather than imaging content.

Figure 3 is consistent with this pattern. Both models show an increase in $\hat{P}(\text{MDD})$ on true MDD patients under conditions that include the MRI prompt (C2, C4), with a symmetric decrease in $\hat{P}(\text{Control})$. The upward shift in $\hat{P}(\text{MDD})$ also appears on the Control cohort, where it should not if the model were responding to genuine diagnostic signals, consistent with priming rather than diagnostic discrimination. We characterize this trigger in detail in Section 7.

## 6 Preliminary Expert Case Study of Reasoning Traces

To assess whether the large F1 gains from multimodal context reflect genuine clinical understanding, we conducted a preliminary human expert case study of reasoning traces generated by `Qwen2.5-VL-3B`, the model exhibiting the most dramatic performance shift.

**Rating protocol** We sampled 12 predictions per class (MDD vs. Control) from the FOR2107 dataset under two conditions: the baseline TEXT(CSV), which receives only the serialized clinical record, and the full multimodal condition TEXT(CSV, PARCEL) + PROMPT(MRI) + PLOT(MRI), which additionally incorporates brain parcellation text and a `nilearn`-generated MRI visualization, yielding 24 reasoning traces per condition (48 total). A clinical psychologist evaluated each trace on three criteria (Faithfulness, Clinical Accuracy & Safety, Diagnostic Reasoning) on a 4-point Likert scale. We refer to the full rubric scores in Appendix I.

| Grp. | Cond., Class | Faith. ↑ | Clin. Acc. ↑ | Reasoning ↑ |
|---|---|---|---|---|
| 0 | C1, Control | $2.17 \pm 1.09$ | $2.25 \pm 0.90$ | $2.46 \pm 1.10$ |
| 1 | C1, MDD | $\underline{2.08} \pm 0.88$ | $\underline{2.17} \pm 0.92$ | $\underline{2.17} \pm 0.92$ |
| 2 | C4, Control | $2.79 \pm 0.98$ | $\mathbf{2.63} \pm 0.82$ | $\mathbf{2.96} \pm 0.91$ |
| 3 | C4, MDD | $\mathbf{2.83} \pm 0.92$ | $2.46 \pm 0.83$ | $2.83 \pm 0.87$ |

Table 4: Expert scores (Mean ± Std) of `Qwen2.5-VL-3B` reasoning rated by a clinical psychologist on a 4-point Likert scale ($n = 12$ per group, FOR2107). **Bold** indicates the best score and underlined indicates the worst score in each metric column. See Table 1 for notation description.

**Results** Table 4 reports results for each criterion across the four groups. Every group scores below 3.0 on average on all three criteria, and faithfulness in particular shows fabrication throughout, with the model adding unverified clinical details even when multimodal context is available. Traces under the full multimodal condition (Groups 2 and 3) do score higher than the baseline (Groups 0 and 1) on all three criteria, and Control predictions under multimodal input (Group 2) reach the highest faithfulness ($2.79 \pm 0.98$) and diagnostic reasoning ($2.96 \pm 0.91$). However, even these stay below

| Cond. | Input | FOR2107 | | OASIS-3 | |
|---|---|---|---|---|---|
| | | Qwen2.5-VL-3B | Ministral-3-3B | Qwen2.5-VL-3B | Ministral-3-3B |
| C1 | TEXT(CSV) | 0.153 | 0.064 | 0.262 | 0.504 |
| C2 | TEXT(CSV) + PROMPT(MRI) | **0.728** | **0.480** | **0.589** | 0.541 |
| C2$_\dagger$ | TEXT(CSV) + PROMPT(FMRI) | 0.702 | 0.361 | 0.379 | **0.569** |
| C2$_\ddagger$ | TEXT(CSV) + PROMPT(WEATHER) | 0.056 | 0.031 | 0.148 | 0.425 |

Table 5: F1 under preamble ablation: factually false fMRI mention and semantically irrelevant weather preamble, on Qwen2.5-VL-3B and Ministral-3-3B across both cohorts. **Bold** indicates best score per model per cohort.

the threshold. The gains in F1 from the multimodal context are therefore not matched by a better evidence-based reasoning, also suggesting that part of these gains comes from priming-induced confidence rather than genuine use of clinical evidence.

## 7 The Preamble Trigger

While modality collapse is well documented (Sim et al., 2025), to our knowledge, prior work has not yet localized which part of the textual input drives the collapse in a clinical setting. Section 5 shows that prompt framing accounts for the majority of the observed shift. We now ask what class of inputs is sufficient to trigger this shift, and whether it is specific to the exact preamble wording used in our pipeline. We first characterize the trigger on Qwen2.5-VL-3B from FOR2107, then validate the findings on Ministral-3-3B and OASIS-3 via an ablation. For formal details we refer to Appendix K.

**Phrase probe** We construct candidate preamble phrases spanning five semantic categories: *MRI/neuroimaging, general clinical, authoritative framing, neutral, structural/format* (full list in Table 12 of Appendix K). Each phrase is used as a probe replacement for the original PROMPT(MRI) on FOR2107 MDD patients with Qwen2.5-VL-3B, and we measure how much each phrase shifts $\hat{P}(\text{MDD})$ relative to the TEXT(CSV) baseline. We additionally compute the cosine similarity of each phrase's induced hidden-state shift to that of the original MRI preamble. Figure 4 plots both quantities jointly.

Neuroimaging phrases sit in the top-right quadrant (Table 12), but so does a general clinical phrase that *announces an available result*, whereas phrases that *describe* a clinical assessment (a diagnosis being established, a specialist evaluation) and format/neutral phrases fall at or below zero on both axes. The trigger is thus tied to an *announcement register*, the claim that some result is available, of

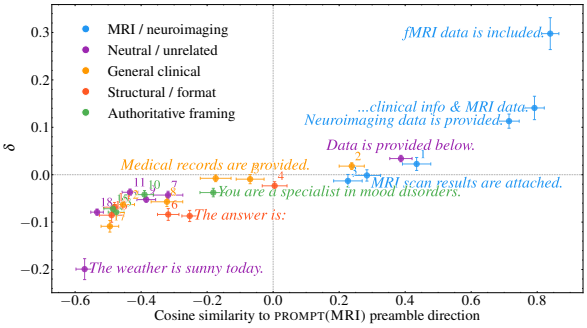



Figure 4: Cosine similarity to the preamble direction (vertical) vs. the shift in mean $\hat{P}(\text{MDD})$ relative to the TEXT(CSV) baseline (horizontal), for candidate phrases across semantic categories, evaluated on Qwen2.5-VL-3B over the FOR2107 cohort. Phrases in the top-right quadrant activate the same effect as TEXT(CSV)+PROMPT(MRI) without providing any imaging data.

which MRI availability is the strongest case, rather than to clinical authority or to descriptions of clinical assessment.

**False-modality ablation** Furthermore, to test whether the trigger is purely surface-level and can be correlationally generalize beyond the model and dataset used in the phrase probe, we use two conditions: a factually false fMRI availability without providing any imaging, and a semantically irrelevant weather context preamble serving as a non-clinical null condition. We evaluate both on Qwen2.5-VL-3B and Ministral-3-3B across both FOR2107 and OASIS-3. Table 5 shows that on FOR2107, the fMRI preamble alone achieves F1 = 0.702 and 0.361 for Qwen2.5-VL-3B and Ministral-3-3B respectively, recovering most of the gain observed under the full multimodal condition, while the weather preamble collapses performance to near or below the text-only baseline (0.056 and 0.031). The contrast replicates on OASIS-3, where the weather condition consistently underperforms the fMRI preamble across both models (extended to two further models

| Cond. | Before MPO ↑ | After MPO ↑ |
|---|---|---|
| C1 | 0.153 | 0.459 |
| C2 | **0.728** | **0.496** |
| C4 | **0.728** | 0.477 |

Table 6: F1 of `Qwen2.5-VL-3B` on FOR2107 before and after MPO. Random baseline F1 = 0.52. **Bold** indicates the best score. See Table 1 for notation description.

that exhibit the F1 gain, `InternVL3.5-4B` and `GLM-4.1V-9B`, in Appendix L). This suggests the trigger correlationally generalizes across model families, with the same qualitative contrast (fMRI preamble above weather) holding on OASIS-3.

## 8 Preference Learning Intervention

The model frequently grounded predictions in MRI-derived or brain-parcellation features even when those features were uninformative, a failure mode confirmed by a domain expert via the preliminary case study. To suppress this behavior, we apply alignment via Direct Preference Optimization (DPO) (Rafailov et al., 2023) in its multimodal extension, Mixed Preference Optimization (MPO) (Wang et al., 2024).

**Dataset construction** We construct a preference dataset from OASIS-3 using the outputs of the 12 models described in Section 4. Each sample belongs to one of three input modes: TEXT(CSV), TEXT(CSV) + PROMPT(MRI), TEXT(CSV, PARCEL) + PROMPT(MRI) + PLOT(MRI). For every subject, we define the *chosen* response as the TEXT(CSV) mode output of any model that predicted the correct label, and the *rejected* response as any output, from any mode, whose text explicitly mentions MRI findings, brain parcellation, or related neuroimaging language. Since we define the rejected response by fabricated MRI usage, the objective targets that surface cue by design. The final dataset comprises 17,202 preference pairs balanced across the two target classes: *Cognitive Normal* and *Cognitive Decline* (≈50% each).

**Evaluation and Results** We fine-tune `Qwen2.5-VL-3B` using MPO. Training details are reported in Appendix M. We then evaluate the trained model on FOR2107 to avoid the naive improvements that would result from in-distribution finetuning, since the preference dataset was constructed from OASIS-3 outputs. Table 6 compares pre- and post-alignment F1 on FOR2107. Preference alignment has opposite effects on the two evaluation modes. In the text-only setting, performance improves substantially ($0.153 \rightarrow 0.459$, $+0.306$). In the full multimodal setting, performance drops ($0.728 \rightarrow 0.477$, $-0.251$). The aligned model no longer produces outputs that reference MRI findings or brain parcellation, satisfying the primary objective of this intervention. However, both modes converge below the unaligned multimodal 0.728. The alignment does not lift the weak mode up to the strong one. We hypothesize that the multimodal mode drop admits two accounts.

- **Account 1 (entanglement).** The strong unaligned multimodal performance was driven by the reasoning mode activated by the preamble, identified in our contrastive analysis. MPO suppresses this mode together with the surface MRI-mention feature removed by design.
- **Account 2 (surface suppression).** MPO learned the MRI-suppression objective alone. The post-alignment F1 drop reflects the loss of a useful textual signal rather than the suppression of a deeper inference pathway.

We discuss the implications of both accounts in Appendix M, where we characterize this trade-off as an open problem for preference alignment.

## 9 Conclusion

We present a clinical case of the modality-collapse phenomenon under strong negative controls: two cohorts selected for documented absence of individual-level diagnostic signal from neuroimaging, with trivially discriminative features excluded in consultation with domain experts. Under these conditions, smaller VLMs exhibit substantial F1 gains upon introduction of multimodal context. We trace the trigger of these gains to the textual preamble itself, characterized via a phrase-level probe and a false-modality ablation, and replicated on additional model families. Preliminary expert case study finds faithfulness remains low in every condition examined, and preference alignment suppresses MRI-referencing behavior but collapses performance across all tested conditions rather than selectively removing modality dependence. Our findings call for evaluation protocols that study the causal relationship between input content and model outputs before VLMs are deployed in clinical settings.

## Limitations

Our study is limited to open-weight models up to 72B parameters, evaluated on two binary classification cohorts. Generalization to closed systems, larger scales, and multi-class or richer clinical tasks remains open. All results use deterministic (greedy) decoding for reproducibility, and we do not examine whether the effect persists under sampling-based decoding.

The preamble-induced gain is not uniform across domains. On OASIS-3, where the preamble was not developed, it degrades several models that are already competitive under text-only input, consistent with negative transfer. We therefore describe the effect as an in-domain vulnerability whose direction and magnitude are domain- and model-dependent, rather than a gain that transfers across clinical settings.

Our confidence analysis covers the two models with the highest gain (extended to four in the appendix), so the reported share of the calibration change attributable to prompt framing should be read in that context. We emphasize, however, that the generality of the effect does not rest on this subset: its presence across all twelve models is shown by the swap-image result (Section 4.1) and its consistency across families by the false-modality ablation (Section 7).

The preliminary expert case study uses a single clinical psychologist, since qualified annotators for patient-level FOR2107 traces are scarce. We therefore treat it as a qualitative case study rather than a quantitative evaluation, with absolute scores regarded as indicative and multi-annotator validation left for future work.

The residual-stream preamble direction in Appendix K is a correlational signature recovered on a single model at a single layer chosen by a fixed onset criterion. We do not verify that the same geometry holds across model families, and establishing causal responsibility for the C1→C2 shift would require intervention experiments, which we leave to future work.

Preference alignment via MPO suppresses MRI-referencing behavior but does not produce factually correct traces. Our dataset also cannot separate the two accounts of the post-alignment collapse in Section 8, since MRI-mention status and correctness are correlated by construction. Finally, our probe centers on a single MRI-framing preamble, and the broader space of triggering phrases and the role of tabular serialization remain to be characterized.

## Ethics Statement

This study uses two clinical datasets governed by strict data-use agreements. FOR2107 data access was granted under the consortium's institutional review protocol, which requires formal application and prohibits redistribution to third parties. OASIS-3 is available through a controlled-access application process administered by the Washington University Knight ADRC, subject to a data-use agreement that similarly prohibits redistribution. No individually identifiable patient information is presented in this work. All reported results are aggregate statistics computed over the full cohorts.

The study on reasoning traces was conducted by a qualified clinical psychologist who reviewed only model-generated outputs. No patient-facing clinical decisions were made on the basis of model predictions, and no real patient data was exposed beyond what is already accessible under the relevant data-use agreements.

We emphasize that the VLM outputs analyzed in this study are not suitable for clinical deployment. Models achieve measurable performance gains that nonetheless carry no diagnostic grounding, produce unverified neuroimaging-grounded justifications, and cannot be straightforwardly corrected through preference alignment without collapsing overall predictive capacity. The gap between surface classification performance and evidence-grounded inference is precisely the vulnerability this work seeks to expose. Any future application of VLMs in clinical psychiatry or neurology must incorporate rigorous prospective validation by domain experts, with evaluation protocols that explicitly probe the causal relationship between input evidence and model outputs, before any patient-facing use is considered.

## References


Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3.

Alberto Mario Ceballos-Arroyo, Monica Munnangi, Jiuding Sun, Karen Zhang, Jered McInerney, Byron C. Wallace, and Silvio Amir. 2024. Open (clinical)

LLMs are sensitive to instruction phrasings. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 50–71, Bangkok, Thailand. Association for Computational Linguistics.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. 2021. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619.

Marie E. Galteau, Margaret Broadwater, Yi Chen, Gabriel Desrosiers-Gregoire, Rita Gil, Johannes Kaesser, Eugene Kim, Pervin Kıryağdı, Henriette Lambers, Yanyan Y. Liu, Xavier López-Gil, Eilidh MacNicol, Parastoo Mohebkhodaei, Ricardo X.N. De Oliveira, Carolina A. Pereira, Henning M. Reimann, Alejandro Rivera-Olvera, Erwan Selingue, Nikoloz Sirmpilatze, and 32 others. 2025a. Activation mapping in multi-center retrospective rat sensory-evoked functional mri datasets using a unified pipeline. *Imaging Neuroscience*, 3:IMAG.a.157.

Marie E. Galteau, Margaret Broadwater, Yi Chen, Gabriel Desrosiers-Gregoire, Rita Gil, Johannes Kaesser, Eugene Kim, Pervin Kıryağdı, Henriette Lambers, Yanyan Y. Liu, Xavier López-Gil, Eilidh MacNicol, Parastoo Mohebkhodaei, Ricardo X.N. De Oliveira, Carolina A. Pereira, Henning M. Reimann, Alejandro Rivera-Olvera, Erwan Selingue, Nikoloz Sirmpilatze, and 32 others. 2025b. Activation mapping in multi-center retrospective rat sensory-evoked functional mri datasets using a unified pipeline. *Imaging Neuroscience*, 3:IMAG.a.157.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Aditya Gulati, Moreno D'Incà, Nicu Sebe, Bruno Lepri, and Nuria Oliver. 2025. Beauty and the bias: Exploring the impact of attractiveness on multimodal large language models. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2):1154–1168.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.

Phillip Howard, Kathleen C. Fraser, Anahita Bhiwandiwalla, and Svetlana Kiritchenko. 2025. Uncovering bias in large vision-language models at scale with counterfactuals. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5946–5991, Albuquerque, New Mexico. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yoichi Ishibashi, Danushka Bollegala, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Evaluating the robustness of discrete prompts. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2373–2384, Dubrovnik, Croatia. Association for Computational Linguistics.

Mohamed Insaf Ismithdeen, Muhammad Uzair Khattak, and Salman Khan. 2025. Promptception: How sensitive are large multimodal models to prompts? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23950–23985, Suzhou, China. Association for Computational Linguistics.

Adrian Javaloy, Maryam Meghdadi, and Isabel Valera. 2022. Mitigating modality collapse in multimodal VAEs via impartial optimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9938–9964. PMLR.

Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. In *Advances in Neural Information Processing Systems*.

Shreya Kapoor and Bernhard Egger. 2025. Computer graphics from a neuroscientist's perspective. In *Second Workshop on Representational Alignment at ICLR 2025*.

Tilo Kircher, Markus Wöhr, Igor Nenadic, Rainer Schwarting, Gerhard Schratt, Judith Alferink, Carsten Culmsee, Holger Garn, Tim Hahn, Bertram Müller-Myhsok, Astrid Dempfle, Maik Hahmann, Andreas Jansen, Petra Pfefferle, Harald Renz, Marcella Rietschel, Stephanie H Witt, Markus Nöthen,

Axel Krug, and Udo Dannlowski. 2019. Neurobiology of the major psychoses: a translational perspective on brain structure and function—the FOR2107 consortium. *European Archives of Psychiatry and Clinical Neuroscience*, 269(8):949–962.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand. Association for Computational Linguistics.

Pamela J. LaMontagne, Tammie LS. Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G. Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. 2019. Oasis-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medRxiv*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, and 1 others. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.

Sheng Lu, Hendrik Schuff, and Iryna Gurevych. 2024. How are prompts different in terms of sensitivity? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5833–5856, Mexico City, Mexico. Association for Computational Linguistics.

David E. Meyer and Roger W. Schvaneveldt. 1971. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, 90 2:227–34.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Letitia Parcalabescu and Anette Frank. 2023. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4032–4059, Toronto, Canada. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

R. Jarrett Rushmore, Kyle Sunderland, Holly Carrington, Justine Chen, Michael Halle, Andras Lasso, G. Papadimitriou, N. Prunier, Elizabeth Rizzoni, Brynn Vessey, Peter Wilson-Braun, Yogesh Rathi, Marek Kubicki, Sylvain Bouix, Edward Yeterian, and Nikos Makris. 2022. Anatomically curated segmentation of human subcortical structures in high resolution magnetic resonance imaging: An open science approach. *Frontiers in Neuroanatomy*, Volume 16 - 2022.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2026. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Yan Shu, Hangui Lin, Yexin Liu, Yan Zhang, Gangyan Zeng, Yan Li, Yu ZHOU, Ser-Nam Lim, Harry Yang, and Nicu Sebe. 2025. When semantics mislead vision: Mitigating large multimodal models hallucinations in scene text spotting and understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. 2025. Can VLMs actually see and read? a survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470, Vienna, Austria. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Doan Nam Long Vu, Rui Tan, Lena Moench, Svenja Jule Francke, Daniel Woiwod, Florian Thomas-Odenthal, Sanna Stroth, Tilo Kircher, Christiane Hermann, Udo Dannlowski, and 1 others. 2025.

Roleplaying with structure: Synthetic therapist-client conversation generation from questionnaires. *arXiv preprint arXiv:2510.25384*.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and 1 others. 2024. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Nils R Winter, Julian Blanke, Ramona Leenings, Jan Ernsting, Lukas Fisch, Kelvin Sarink, Carlotta Barkhau, Daniel Emden, Katharina Thiel, Kira Flinkenflügel, Alexandra Winter, Janik Goltermann, Susanne Meinert, Katharina Dohm, Jonathan Repple, Marius Gruber, Elisabeth J Leehr, Nils Opel, Dominik Grotegerd, and 26 others. 2024. A systematic evaluation of machine Learning-Based biomarkers for major depressive disorder. *JAMA Psychiatry*, 81(4):386–395.

Nils R Winter, Ramona Leenings, Jan Ernsting, Kelvin Sarink, Lukas Fisch, Daniel Emden, Julian Blanke, Janik Goltermann, Nils Opel, Carlotta Barkhau, Susanne Meinert, Katharina Dohm, Jonathan Repple, Marco Mauritz, Marius Gruber, Elisabeth J Leehr, Dominik Grotegerd, Ronny Redlich, Andreas Jansen, and 12 others. 2022. Quantifying deviations of brain structure and function in major depressive disorder across neuroimaging modalities. *JAMA Psychiatry*, 79(9):879–888.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Yiheng Xu, Peng Wang, Hang Zhang, Pengfei Wang, Shuai Bai, Shijie Wang, Junyang Lin, Tianbao Xie, Yuanzhi Zhu, Zhibo Yang, Wei Ding, Xi Zhang, Jianqiang Wan, Jun Tang, Haiyang Xu, Jiabo Ye, Keqin Chen, Xuejing Liu, Jialin Wang, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Daiki Yoshida, Haruki Sakajo, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2025. Visual priming effect on large-scale vision language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1385–1395, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Chi Zhang, Wenxuan Ding, Jiale Liu, Mingrui Wu, Qingyun Wu, and Ray Mooney. 2026. Do images speak louder than words? investigating the effect of textual misinformation in VLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6872–6895, Rabat, Morocco. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11991–12011, Bangkok, Thailand. Association for Computational Linguistics.

Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Wang, Qi Wu, Miguel Eckstein, and William Yang Wang. 2022. Diagnosing vision-and-language navigation: What really matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5981–5993, Seattle, United States. Association for Computational Linguistics.

## A Prompt

Prompts that are used in our paper (Figure 7, 8 for FOR2107), (Figure 9, 10 for OASIS-3). Label answers from every condition and model are extracted at a 100% parse rate using the fixed JSON prefix `{"category": "` (See Table 7 for each model prefix IDs)

| Model | Prefix token IDs | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| GLM-4.1V-9B | 73022 | 2236 | 198 | 515 | 220 | 330 | 5471 | 788 | 330 |
| GLM-4.6V-Flash | 73022 | 2236 | 198 | 515 | 220 | 330 | 5471 | 788 | 330 |
| LLaVA-OV-1.5-4B | 73594 | 2236 | 198 | 515 | 220 | 330 | 5471 | 788 | 330 |
| Ministral-3-3B | 1975 | 8353 | 1010 | 2030 | 1032 | 1429 | 17278 | 2811 | 1429 |
| Ministral-3-14B | 1975 | 8353 | 1010 | 2030 | 1032 | 1429 | 17278 | 2811 | 1429 |
| Qwen2.5-VL-3B | 73594 | 2236 | 198 | 515 | 220 | 330 | 5471 | 788 | 330 |
| Qwen2.5-VL-32B | 73594 | 2236 | 198 | 515 | 220 | 330 | 5471 | 788 | 330 |
| Qwen2.5-VL-72B | 73594 | 2236 | 198 | 515 | 220 | 330 | 5471 | 788 | 330 |
| Qwen3-VL-2B | 73594 | 2236 | 198 | 515 | 220 | 330 | 5471 | 788 | 330 |
| Qwen3-VL-32B | 73594 | 2236 | 198 | 515 | 220 | 330 | 5471 | 788 | 330 |
| InternVL3.5-4B | 73594 | 2236 | 198 | 515 | 220 | 330 | 5471 | 788 | 330 |
| InternVL3.5-14B | 73594 | 2236 | 198 | 515 | 220 | 330 | 5471 | 788 | 330 |
| medgemma-4b-it | 2717 | 3723 | 107 | 236782 | 107 | 138 | 236775 | 10618 | 1083 |
| medgemma-27b-it | 2717 | 3723 | 107 | 236782 | 107 | 138 | 236775 | 10618 | 1083 |

Table 7: Token IDs of the JSON prefix for all evaluated models in the paper.

| Abbreviation | Checkpoint (Link to HF) |
|---|---|
| GLM-4.1V-9B | zai-org/GLM-4.1V-9B-Thinking |
| GLM-4.6V-Flash | zai-org/GLM-4.6V-Flash |
| LLaVA-OV-1.5-4B | lmms-lab/LLaVA-OneVision-1.5-4B-Instruct |
| Ministral-3-3B | mistralai/Ministral-3-3B-Instruct-2512-BF16 |
| Ministral-3-14B | mistralai/Ministral-3-14B-Instruct-2512-BF16 |
| Qwen2.5-VL-3B | Qwen/Qwen2.5-VL-3B-Instruct |
| Qwen2.5-VL-32B | Qwen/Qwen2.5-VL-32B-Instruct |
| Qwen2.5-VL-72B | Qwen/Qwen2.5-VL-72B-Instruct |
| Qwen3-VL-2B | Qwen/Qwen3-VL-2B-Instruct |
| Qwen3-VL-32B | Qwen/Qwen3-VL-32B-Instruct |
| InternVL3.5-4B | OpenGVLab/InternVL3_5-4B |
| InternVL3.5-14B | OpenGVLab/InternVL3_5-14B |
| *Domain-specialized ablation* | |
| medgemma-4b-it | google/medgemma-4b-it |
| medgemma-27b-it | google/medgemma-27b-it |

Table 8: Abbreviation for models used in this work and Hyperlinked Checkpoint Paths

## B Parcellation in Detail

Parcellation is performed using the Harvard-Oxford probabilistic atlas (Rushmore et al., 2022), which delineates the cerebral cortex into 48 anatomically defined regions of interest (ROIs, e.g., Insular Cortex, Temporal Pole). The atlas was constructed by spatially normalizing manually labeled T1-weighted scans of healthy adults into MNI (Montreal Neurological Institute) space and computing voxelwise label probabilities across subjects. We apply the cortical component of this atlas to extract per-region volumetric measurements, which are then serialized into descriptive text following the same approach as the tabular data.

**MRI visualization** For each scan, we render three orthogonal slices, sagittal, coronal, and axial, annotated with crosshairs at MNI coordinates $(x, y, z)$, hemisphere labels (L/R), voxel contour overlays, and a zero-anchored colormap to enhance tissue contrast.

## C Models in Detail

Table 8 shows the abbreviations and checkpoints of the models we used in this work.

## D FOR2107 Variables

Tables 17, 18, 19, 20, 21, and 22 list the variables we serialize into text. Figure 5 shows the distribution of input tokens in FOR2107, split by Major

Depressive Disorder and Control patients. We use `tiktoken`[5] for tokenization.

## E OASIS-3 Variables

Tables 25, 26, 28, and 27 list the variables we serialize into text. Figure 6 shows the token distribution of input in OASIS-3, split by Cognitive Normal and Cognitive Decline patients. We also use `tiktoken` for tokenization.

## F Random Baseline

To contextualize model performance, we report a random baseline corresponding to a stratified random classifier that predicts each class with a probability equal to its prior. For a binary classification task with class proportions $p$ and $1-p$, the expected weighted F1 score of such a classifier is: $F1_{\text{random}} = p^2 + (1-p)^2$, where $p$ is the proportion of the minority class.

## G Out-of-Distribution Images for C5

Condition C5 replaces the subject-specific MRI plot with an out-of-distribution image while retaining all other input components. Two image variants are used across the experiments, shown in Figure 11: an unrelated natural photograph (a dog) and a stylized non-clinical brain rendering. Neither image carries diagnostic information about the patient.

## H Full Results on FOR2107 and OASIS-3

### H.1 FOR2107 - Per-Condition Results

Table 13 reports F1, Precision, Recall, and Accuracy for all twelve models across C1-C5. The pattern in the main text holds throughout: under C1 several small models fail to exceed the random baseline (`InternVL3.5-4B` 0.000, `Ministral-3-3B` 0.064, `Qwen2.5-VL-3B` 0.153), and adding the MRI preamble (C2) produces large gains for exactly these models while leaving strong models nearly unchanged. Adding the actual image (C3) and parcellation features (C4) yields no consistent further improvement, and the counterfactual swap (C5) leaves performance close to C4 (e.g. `Qwen2.5-VL-72B` 0.849 in both), confirming that the gains track the preamble framing rather than image content.

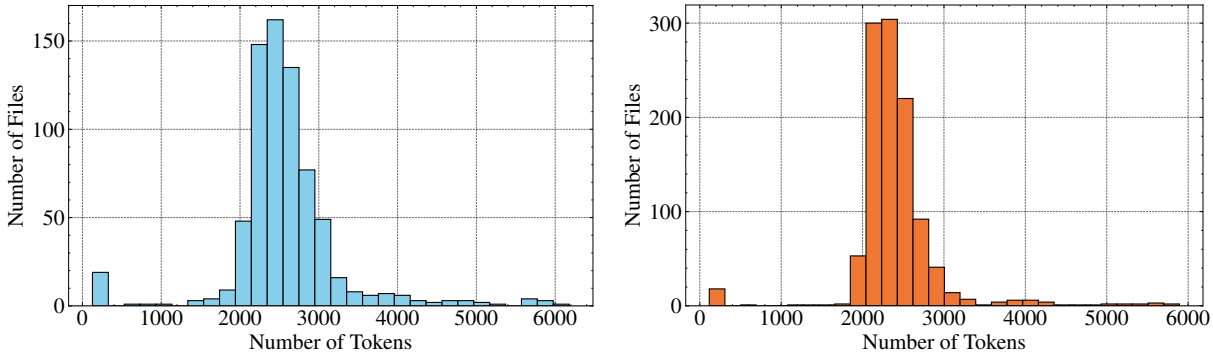
[5] https://github.com/openai/tiktoken

### H.2 OASIS-3 - Per-Condition Results

Table 14 reports F1, Precision, Recall, and Accuracy for all twelve models across C1-C5 on OASIS-3, a more class-balanced cohort in a different clinical domain (cognitive decline vs. depression). The picture here is more mixed than on FOR2107. Several models that are already competitive under C1 (e.g. `InternVL3.5-14B`, 0.773) *degrade* when the MRI preamble is added, consistent with negative transfer: the preamble was developed on the FOR2107 domain and disrupts well-calibrated text-only behavior out of domain. Unlike on FOR2107, the actual image (C3) tends to help more than the preamble alone, and parcellation features (C4) again add little. Most relevant to our claim, the counterfactual swap (C5) again closely mirrors C4 (e.g. `Qwen2.5-VL-72B` 0.778 in both), so where multimodal gains do appear they still do not depend on genuine image content. We read OASIS-3 as further evidence that apparent multimodal gains track preamble framing rather than image content, in a regime where the preamble itself transfers poorly.

**Ablation `medgemma` models** `medgemma` (Sellergren et al., 2026) is Google's medical-specialized open-weight VLM, with a SigLIP image encoder pretrained on de-identified medical imagery (including radiology) and an LLM component trained on medical text and clinical records. We include it to test whether pretraining on medical data makes a VLM immune to the preamble-induced shift.

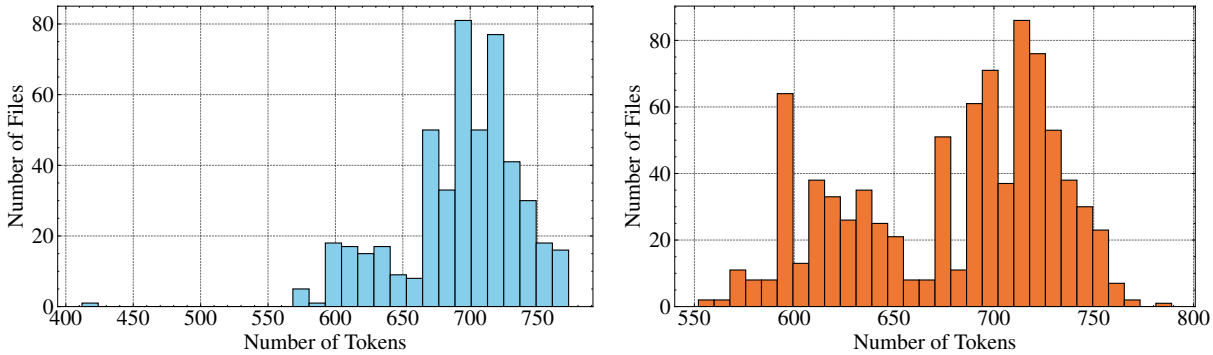
The two `medgemma` model sizes fail in opposite directions across the two datasets. On FOR2107, `medgemma-27b-it` is the stronger model across all conditions. F1 rises from 0.467 to 0.582 when the MRI preamble is added with no imaging data, lifting the model from below to above the random baseline. `medgemma-4b-it` is non-functional on this task, with near-zero recall across C3-C5. On OASIS-3 the roles invert: `medgemma-4b-it` maintains a strong text-only baseline (C1 F1 $=$ 0.739) and `medgemma-27b-it` collapses to a near-degenerate negative-prediction strategy under C1 and C2 (F1 $\approx$ 0.07, P $=$ 1.000, R $\approx$ 0.04). Crucially, both OASIS-3 models produce *byte-identical* outputs under C1 and C2. This points to a different failure mode: `medgemma`'s predictions appear driven by the presence and form of the image input rather than by the framing of the prompt. Neither model size approaches the text-only performance of the strongest general-purpose VLMs




(a) Major Depressive Disorder Patient

(b) Control Patient

Figure 5: Distribution of input tokens in FOR2107 dataset




(a) Cognitive Decline Patient

(b) Cognitive Normal Patient

Figure 6: Distribution of input tokens in OASIS-3 dataset

**Prompt**

You are given patient clinical information
Classify the patient into one of the following categories:
- Major Depressive Disorder
- Control (no disorder detected)
Return your answer as a JSON object with two fields:
- "category": the chosen category (exactly one of the two above)
- "explanation": a short reasoning for the choice

Figure 7: TEXT(CSV) prompt for FOR2107

**Prompt**

You are given patient clinical information **and their MRI data (brain parcellation volume, visualization of brain regions)**
Classify the patient into one of the following categories:
- Major Depressive Disorder
- Control (no disorder detected)
Return your answer as a JSON object with two fields:
- "category": the chosen category (exactly one of the two above)
- "explanation": a short reasoning for the choice

Figure 8: TEXT(CSV) + PROMPT(MRI) prompt for FOR2107. **Bold text** denotes the PROMPT(MRI) component that cause performance boost.

in Table 14, and the cross-dataset role reversal between the 4B and 27B checkpoints indicates that, for medgemma, medical-domain pretraining did not by itself yield reliable clinical reasoning across both cohorts.

### H.3 Expert Consultation

The exclusion of trivially discriminative features described in Section 3 was conducted in consultation with domain experts in clinical psychology. The consulted experts hold doctoral degrees in clinical psychology or a closely related field, with the majority being postdoctoral researchers and a minority advanced doctoral candidates, each with active research experience on the relevant clinical populations. Their identities are withheld for the purpose of double-blind review.

| Dataset | Cond. | medgemma-4b-it | | | | medgemma-27b-it | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | F1 | P | R | ACC | F1 | P | R | ACC |
| FOR2107 | C1 | 0.130 | 0.891 | 0.070 | 0.629 | **0.467** | 0.982 | 0.307 | 0.724 |
| | C2 | 0.100 | 0.902 | 0.053 | 0.623 | **0.582** | 0.914 | 0.427 | 0.757 |
| | C3 | 0.003 | 1.000 | 0.001 | 0.605 | **0.625** | 0.925 | 0.472 | 0.776 |
| | C4 | 0.014 | 1.000 | 0.007 | 0.607 | **0.386** | 0.966 | 0.241 | 0.696 |
| | C5 | 0.003 | 1.000 | 0.001 | 0.605 | **0.322** | 0.951 | 0.194 | 0.677 |
| OASIS-3 | C1 | **0.739** | 0.876 | 0.639 | 0.835 | 0.071 | 1.000 | 0.037 | 0.649 |
| | C2 | **0.739** | 0.876 | 0.639 | 0.835 | 0.071 | 1.000 | 0.037 | 0.649 |
| | C3 | 0.413 | 0.774 | 0.281 | 0.708 | **0.606** | 0.503 | 0.760 | 0.639 |
| | C4 | **0.591** | 0.445 | 0.877 | 0.556 | 0.443 | 0.740 | 0.316 | 0.710 |
| | C5 | **0.576** | 0.450 | 0.801 | 0.570 | 0.400 | 0.798 | 0.267 | 0.708 |

Table 9: `medgemma` models ablation on FOR2107 and OASIS-3: F1, Precision (P), Recall (R), and Accuracy (ACC) across the five input conditions. **Bold** marks the higher F1 between the two model sizes per row. See Table 1 for condition notation.

**Prompt**

You are given patient clinical information
Classify the patient into one of the following categories:
- Cognitive Normal
- Cognitive Decline
Return your answer as a JSON object with two fields:
- "category": the chosen category (exactly one of the two above)
- "explanation": a short reasoning for the choice

Figure 9: TEXT(CSV) prompt for OASIS-3

**Prompt**

You are given patient clinical information **and their MRI data (brain parcellation volume, visualization of brain regions)**
Classify the patient into one of the following categories:
- Cognitive Normal
- Cognitive Decline
Return your answer as a JSON object with two fields:
- "category": the chosen category (exactly one of the two above)
- "explanation": a short reasoning for the choice

Figure 10: TEXT(CSV) + PROMPT(MRI) prompt for OASIS-3. **Bold text** denotes the PROMPT(MRI) component that cause performance boost.

# I Preliminary Case Study - Score Rubric in Detail

- **Faithfulness**: whether the output strictly adheres to the input context without adding unverified information (1 = severe fabrication, 4 = perfectly faithful).
- **Clinical Accuracy & Safety**: whether the conclusions are correct, safe, and aligned with clinical standards (1 = completely inaccurate or unsafe, 4 = perfectly accurate and safe).
- **Diagnostic Reasoning**: whether the reasoning behind the diagnosis is clearly and logically explained (1 = poor, 4 = excellent).

We adopt a mean of 3.0 as a conservative reference for an acceptable level: it is the boundary between the problematic half of the scale (1-2) and the acceptable half (3-4). We use it only as a descriptive anchor for reading the scores, not as a validated clinical cutoff.

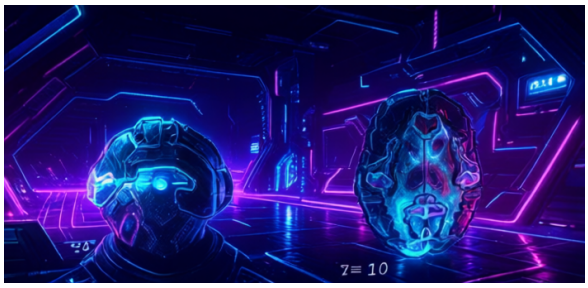
# J Confidence Estimation in Detail

To quantify model confidence during inference, we extract per-token probability from the decoder using greedy decoding with `output_scores=True`, which provides the full vocabulary distribution at each generation step. This formulation follows label-token probability extraction approaches (Zhao et al., 2021; Geng et al., 2024) and combined with the generated prefix differs across conditions.

**Joint formulation.** For each patient $x$ and class $\ell \in \{\text{MDD}, \text{Control}\}$, we consider the joint probability under the model of the generated prefix prefix together with the label token $t^\ell$:

$$P(\text{prefix}, t^\ell \mid x) = P(\text{prefix} \mid x) \cdot P(t^\ell \mid \text{prefix}, x), \quad (1)$$

(a) Natural photograph (dog).




(b) Stylized non-clinical brain rendering.

Figure 11: Out-of-distribution images used in condition C5. Both replace the subject-specific MRI plot while the rest of the multimodal preamble remains unchanged.

where

$$P(\text{prefix} \mid x) = \prod_{j=1}^{|\text{prefix}|} P(\text{prefix}_j \mid \text{prefix}_{<j}, x)$$

is the autoregressive prefix probability. The normalized confidence score is obtained by renormalizing across the two label classes:

$$\hat{P}_c(\ell \mid x) = \frac{P(\text{prefix}, t^{\ell} \mid x)}{\sum_{\ell'} P(\text{prefix}, t^{\ell'} \mid x) + \epsilon}, \quad (2)$$

where $\epsilon = 10^{-12}$ for numerical stability and $c$ denotes the input condition. The predicted class is $\hat{y} = \arg\max_{\ell} \hat{P}_c(\ell \mid x)$. The score $\hat{P}_c(\ell \mid x)$ is bounded in $[0, 1]$ and expresses the model's relative preference between the two labels within a given (patient, condition) pair. All cross-condition comparisons (Eq. 3) are made on this normalized quantity rather than on raw joint probabilities.

**Per-patient confidence shift.** We use $\hat{P}_c(\ell \mid x)$ as the basis for a contrastive analysis designed to disentangle the respective contributions of prompt framing and actual MRI content. Specifically, we evaluate each patient under three conditions and define the per-patient confidence shift as:

$$\delta^{\text{cond}}_{\leftarrow \text{base}}(x_i) = \hat{P}_{\text{cond}}(\ell \mid x_i) - \hat{P}_{\text{base}}(\ell \mid x_i), \quad (3)$$

where $\ell = \text{MDD}$. A positive $\delta$ indicates that the condition increases the model's confidence in the MDD label relative to the baseline.

### J.1 More Models Confidence Estimation

Table 10 reports ECE and Brier for `InternVL3.5-4B` and `GLM-4.1V-9B` on FOR2107 across C1, C2, and C4, complementing the two-model analysis in Section 5. Both models reproduce the central calibration pattern reported there: the C1→C2 transition, in which no imaging content is provided, accounts for the bulk of the calibration improvement. For `InternVL3.5-4B`, 79.5% of the total C1→C4 ECE reduction is realized at C2; for `GLM-4.1V-9B` we observe a similar trend, as C4 in fact has marginally worse calibration than C2 (ECE 0.204 vs. 0.196). Figure 12 reproduces the pattern observed for the original two models: $\hat{P}(\text{MDD})$ increases under C2 on both the MDD and Control cohorts, with a mirror-symmetric decrease in $\hat{P}(\text{Control})$. The upward shift in $\hat{P}(\text{MDD})$ also appears on the Control cohort, where genuine diagnostic signal would not produce it. These results are consistent with the effect being a structural property of how VLMs process domain-specific multimodal preambles.

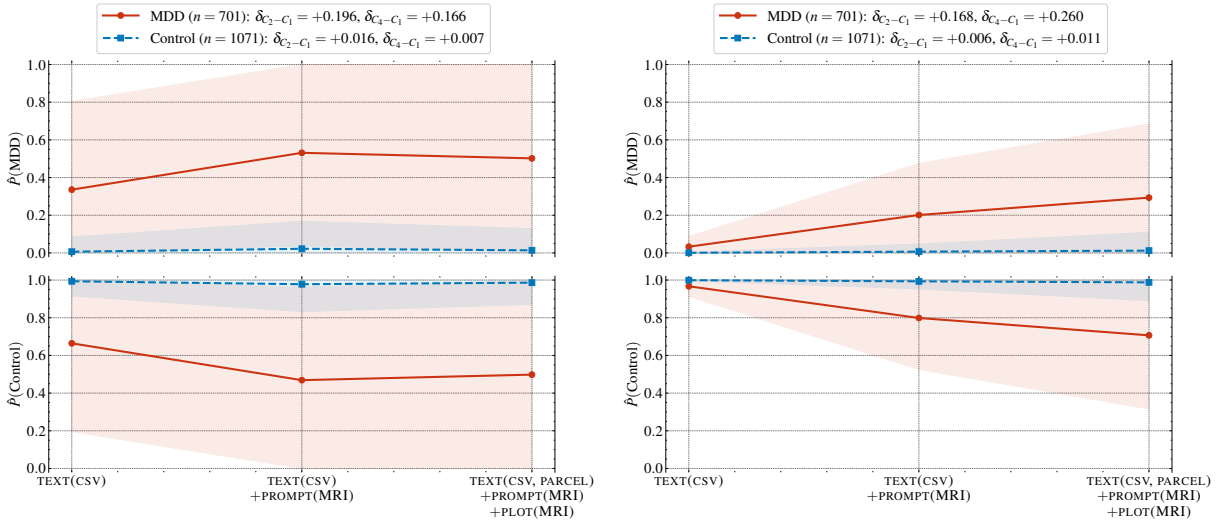
| Model | Cond. | $\text{ECE}_{15}$ | Brier |
|---|---|---|---|
| `InternVL3.5-4B` | C1 | 0.382 | 0.371 |
| | C2 | 0.289 | 0.283 |
| | C4 | **0.265** | **0.264** |
| `GLM-4.1V-9B` | C1 | 0.265 | 0.265 |
| | C2 | **0.196** | **0.196** |
| | C4 | 0.204 | 0.204 |

Table 10: Confidence estimation on FOR2107 for `InternVL3.5-4B` and `GLM-4.1V-9B`, extending the analysis of Section 5 (Table 3). Computed over all 1,772 patients (MDD + Control combined). **Bold** denotes better score.

## K Preamble Trigger: Formal Detail

### K.1 Preamble Direction

Let $\mathcal{M}$ be a VLM with $L$ layers and hidden dimension $d$. Let $\mathbf{h}_c^{(l)}(x) \in \mathbb{R}^d$ denote the residual-stream hidden state at layer $l$, extracted at the label token decoding step for patient record $x$ under input condition $c$. We denote the target positive class as $\ell^+$ (e.g. MDD for FOR2107). For




(a) GLM-4.1V-9B

(b) InternVL3.5-4B

Figure 12: FOR2107 per-cohort mean $\hat{P}(\text{MDD})$ (top) and $\hat{P}(\text{Control})$ (bottom) across the three input conditions. Both cohorts exhibit an upward shift in $\hat{P}(\text{MDD})$, with a mirror-symmetric downward shift in $\hat{P}(\text{Control})$.

brevity we refer to the two main-text conditions of interest as $c_{\text{base}}$ (C1: TEXT(CSV)) and $c_{\text{cond}}$ (C2: TEXT(CSV)+PROMPT(MRI)). Given a set of $N$ patient records $\{x_i\}_{i=1}^{N}$, we define the **preamble direction** at layer $l^*$ as:

$$\mathbf{d} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{h}_{c_{\text{cond}}}^{(l^*)}(x_i) - \mathbf{h}_{c_{\text{base}}}^{(l^*)}(x_i)\right), \quad (4)$$

with unit-normalized form $\mathbf{u} = \mathbf{d}/\|\mathbf{d}\|_2$. Intuitively, $\mathbf{u}$ captures the direction in residual-stream space that the MRI preamble injects to shift the model's final-layer classification routing toward $\ell^+$. The target layer $l^*$ is selected per model as the layer immediately preceding the point at which the label first emerges under a logit-lens sweep, a fixed criterion rather than one tuned to maximize the measured effect.

### K.2 Phrase Probe Inventory

Table 12 lists the full inventory of phrase probes with their scores. We applied a logit-lens sweep[6] over all 36 layers of Qwen2.5-VL-3B for a subset of FOR2107 participants, reading the label at each layer by projecting the residual stream through the unembedding. As shown in Figure 13, the label carries negligible weight through the early and middle layers and emerges only near the final layers (onset $\approx$ layer 32), with the preamble conditions sitting above the C1 baseline past this onset. We therefore extract residual-stream activations at the layer immediately preceding the onset, by a fixed criterion rather than one tuned to maximize the effect. We treat the resulting direction as exploratory and correlational: it characterizes where the preamble's influence becomes visible in the residual stream, not that this direction causally drives the classification shift.

[6] https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens

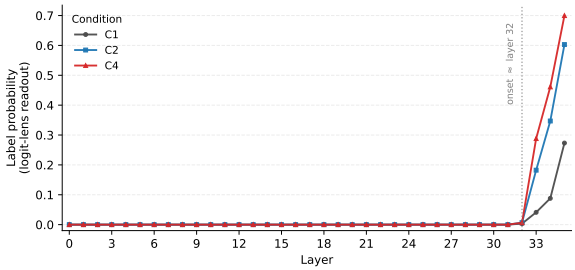



Figure 13: Logit-lens trajectory across all 36 layers of Qwen2.5-VL-3B on FOR2107, under conditions C1, C2, and C4. The label carries negligible weight through the early and middle layers and emerges only in the final layers (onset $\approx$ layer 32). Past this onset the preamble conditions (C2, C4) sit above the C1 baseline, consistent with the preamble shifting weight toward the target label. We extract at the layer immediately preceding the onset.

## L Extended Preamble Ablation

We replicate the preamble ablation on two additional model families, InternVL3.5-4B and GLM-4.1V-9B. Table 11 reports F1 across the four preamble conditions on both cohorts. On

| Cond. | Input | FOR2107 | | OASIS-3 | |
|---|---|---|---|---|---|
| | | InternVL3.5-4B | GLM-4.1V-9B | InternVL3.5-4B | GLM-4.1V-9B |
| C1 | TEXT(CSV) | 0.000 | 0.521 | 0.616 | 0.278 |
| C2 | TEXT(CSV) + PROMPT(MRI) | **0.255** | 0.640 | 0.636 | 0.397 |
| C2† | TEXT(CSV) + PROMPT(FMRI) | 0.186 | **0.661** | **0.652** | **0.527** |
| C2‡ | TEXT(CSV) + PROMPT(WEATHER) | 0.000 | 0.512 | 0.594 | 0.339 |

Table 11: Extended preamble ablation on InternVL3.5-4B and GLM-4.1V-9B, replicating the design of Table 5. **Bold** indicates best score per model per cohort.

FOR2107, InternVL3.5-4B scores 0.000 at C1 and 0.255 under the MRI preamble; GLM-4.1V-9B scores 0.521 at C1 and 0.661 under the fMRI preamble. On OASIS-3, InternVL3.5-4B scores 0.616 at C1 and 0.652 under the fMRI preamble; GLM-4.1V-9B scores 0.278 at C1 and 0.527 under the fMRI preamble. The fMRI preamble improves F1 over the C1 baseline in all four cells, while the weather preamble lowers it except for a small lift on GLM-4.1V-9B on OASIS-3 (+0.061). As expected, given their different baseline capabilities, the absolute gains vary across models. However, the direction of the effect across families is the relevant signal.

## M Preference Learning in Detail

We use MPO for preference finetuning which combines three complementary loss signals:

$$\mathcal{L}_{\text{MPO}} = w_{\sigma}\,\mathcal{L}_{\text{sigmoid}} + w_{\text{bco}}\,\mathcal{L}_{\text{bco_pair}} + w_{\text{sft}}\,\mathcal{L}_{\text{sft}}, \tag{5}$$

where $\mathcal{L}_{\text{sigmoid}}$ is the standard DPO preference loss, $\mathcal{L}_{\text{bco_pair}}$ is a binary quality signal penalizing rejected responses, and $\mathcal{L}_{\text{sft}}$ is a generation loss on the chosen response that prevents degenerate output collapse. Parameter-efficient fine-tuning is performed via Low-Rank Adaptation (LoRA) (Hu et al., 2022), targeting all attention projection and feed-forward layers. Training ran for approximately 12 hours on a single NVIDIA H200 GPU. The LoRA configuration is given in Table 15. Table 16 depicts all hyperparameters used in our training.

**Interpretation of the post-alignment collapse** Distinguishing the two accounts in Section 8 would require a preference dataset in which MRI-mention status and prediction correctness are decorrelated across the chosen and rejected sets. In our current 17,202 pair set the off-diagonal cells {MRI-mention, correct} and {no-MRI-mention, incorrect} are sparsely populated, since correctness and MRI-mention status are correlated by construction in our chosen/rejected definition. Balancing the four cells would require one of: (i) large-scale resampling at elevated decoding temperature, which conflicts with our greedy decoding setup for reproducibility, (ii) controlled editing of existing responses to flip MRI-mention status while preserving correctness, which requires expert annotation at scale, or (iii) synthetic response generation conditioned on each target cell, which suffers from the same annotation bottleneck. We leave a systematic comparison to future work. Regardless of which account is correct, the empirical finding stands. Lightweight preference alignment against MRI-referencing outputs does not selectively remove the dependence on irrelevant modalities while preserving overall performance.

| # | Category | Phrase | cos↑ | $\delta$↑ |
|---|---|---|---|---|
| 1 | • MRI / neuroimaging | *Brain MRI findings are available.* | +0.43 | +0.02 |
| 2 | • General clinical | *Diagnostic results are available.* | +0.24 | +0.02 |
| 3 | • MRI / neuroimaging | *Brain scans have been performed.* | +0.23 | -0.01 |
| 4 | • Structural / format | *Classification result:* | +0.00 | -0.02 |
| 5 | • General clinical | *A clinical diagnosis has been established.* | -0.07 | -0.01 |
| 6 | • Structural / format | *The answer is:* | -0.25 | -0.09 |
| 7 | • Structural / format | *Output: {"category":* | -0.32 | -0.08 |
| 8 | • Neutral / unrelated | *This is a test of the system.* | -0.32 | -0.04 |
| 9 | • General clinical | *Clinical evaluation is complete.* | -0.32 | -0.06 |
| 10 | • Neutral / unrelated | *Answer the following question carefully.* | -0.38 | -0.05 |
| 11 | • Authoritative framing | *As a medical professional, review the following.* | -0.39 | -0.04 |
| 12 | • Neutral / unrelated | *Please process the following information.* | -0.43 | -0.04 |
| 13 | • General clinical | *The patient has been evaluated by a specialist.* | -0.45 | -0.06 |
| 14 | • Authoritative framing | *You have extensive experience in psychiatric diagnosis.* | -0.48 | -0.08 |
| 15 | • Structural / format | *Return your answer as JSON.* | -0.48 | -0.07 |
| 16 | • Authoritative framing | *You are an expert clinical psychiatrist.* | -0.49 | -0.07 |
| 17 | • Structural / format | *Respond only with a JSON object.* | -0.49 | -0.09 |
| 18 | • General clinical | *The patient has been assessed for a psychiatric disorder.* | -0.49 | -0.11 |
| 19 | • Neutral / unrelated | *You are a helpful assistant.* | -0.53 | -0.08 |

Table 12: Numbered phrases from Figure 4 with their scores. cos: cosine similarity between the phrase's induced hidden-state shift and the MRI-preamble direction. $\delta$: change in mean $\hat{P}(\text{MDD})$ on FOR2107 MDD patients relative to the TEXT(CSV) baseline when the phrase replaces the MRI preamble. Higher values on both axes indicate a stronger preamble-like effect.

| Model | F1 | P | R | ACC |
|---|---|---|---|---|
| *C1:* TEXT(CSV) | | | | |
| InternVL3.5-4B | 0.000 | 0.000 | 0.000 | 0.604 |
| InternVL3.5-14B | 0.559 | 0.946 | 0.397 | 0.752 |
| GLM-4.1V-9B | 0.521 | 0.958 | 0.358 | 0.740 |
| GLM-4.6V-Flash | 0.795 | 0.896 | 0.715 | 0.854 |
| LLaVA-OV-1.5-4B | 0.663 | 0.954 | 0.508 | 0.796 |
| Ministral-3-3B | 0.064 | 1.000 | 0.033 | 0.617 |
| Ministral-3-14B | 0.588 | 0.983 | 0.419 | 0.768 |
| Qwen2.5-VL-3B | 0.153 | 1.000 | 0.083 | 0.637 |
| Qwen2.5-VL-32B | 0.749 | 0.900 | 0.642 | 0.830 |
| Qwen2.5-VL-72B | **0.828** | 0.853 | 0.805 | 0.868 |
| Qwen3-VL-2B | 0.546 | 0.869 | 0.398 | 0.738 |
| Qwen3-VL-32B | 0.744 | 0.928 | 0.621 | 0.831 |
| *C2:* TEXT(CSV) + PROMPT(MRI) | | | | |
| InternVL3.5-4B | 0.255 | 0.963 | 0.147 | 0.660 |
| InternVL3.5-14B | 0.729 | 0.884 | 0.621 | 0.818 |
| GLM-4.1V-9B | 0.640 | 0.915 | 0.492 | 0.781 |
| GLM-4.6V-Flash | 0.802 | 0.847 | 0.760 | 0.851 |
| LLaVA-OV-1.5-4B | 0.592 | 0.980 | 0.424 | 0.769 |
| Ministral-3-3B | 0.480 | 0.720 | 0.360 | 0.691 |
| Ministral-3-14B | 0.680 | 0.961 | 0.526 | 0.804 |
| Qwen2.5-VL-3B | 0.728 | 0.876 | 0.623 | 0.816 |
| Qwen2.5-VL-32B | 0.820 | 0.897 | 0.756 | 0.869 |
| Qwen2.5-VL-72B | **0.844** | 0.804 | 0.889 | 0.870 |
| Qwen3-VL-2B | 0.281 | 0.966 | 0.164 | 0.667 |
| Qwen3-VL-32B | 0.823 | 0.900 | 0.758 | 0.871 |
| *C3:* TEXT(CSV) + PROMPT(MRI) + MRI PLOT | | | | |
| InternVL3.5-4B | 0.296 | 0.992 | 0.174 | 0.673 |
| InternVL3.5-14B | 0.540 | 0.981 | 0.372 | 0.749 |
| GLM-4.1V-9B | 0.733 | 0.887 | 0.625 | 0.820 |
| GLM-4.6V-Flash | 0.794 | 0.875 | 0.728 | 0.851 |
| LLaVA-OV-1.5-4B | 0.713 | 0.937 | 0.575 | 0.817 |
| Ministral-3-3B | 0.723 | 0.851 | 0.628 | 0.809 |
| Ministral-3-14B | 0.710 | 0.948 | 0.568 | 0.817 |
| Qwen2.5-VL-3B | 0.680 | 0.878 | 0.555 | 0.794 |
| Qwen2.5-VL-32B | 0.820 | 0.840 | 0.802 | 0.861 |
| Qwen2.5-VL-72B | 0.846 | 0.795 | 0.903 | 0.870 |
| Qwen3-VL-2B | 0.447 | 0.908 | 0.297 | 0.710 |
| Qwen3-VL-32B | **0.846** | 0.827 | 0.866 | 0.875 |

| Model | F1 | P | R | ACC |
|---|---|---|---|---|
| *C4:* TEXT(CSV,PARCEL) + PROMPT(MRI) + MRI PLOT | | | | |
| InternVL3.5-4B | 0.252 | 0.990 | 0.144 | 0.661 |
| InternVL3.5-14B | 0.587 | 0.913 | 0.432 | 0.759 |
| GLM-4.1V-9B | 0.720 | 0.916 | 0.593 | 0.818 |
| GLM-4.6V-Flash | 0.791 | 0.848 | 0.740 | 0.845 |
| LLaVA-OV-1.5-4B | 0.603 | 0.975 | 0.437 | 0.773 |
| Ministral-3-3B | 0.677 | 0.903 | 0.542 | 0.796 |
| Ministral-3-14B | 0.680 | 0.966 | 0.525 | 0.805 |
| Qwen2.5-VL-3B | 0.723 | 0.633 | 0.845 | 0.744 |
| Qwen2.5-VL-32B | 0.779 | 0.896 | 0.689 | 0.845 |
| Qwen2.5-VL-72B | **0.845** | 0.798 | 0.897 | 0.870 |
| Qwen3-VL-2B | 0.347 | 0.974 | 0.211 | 0.686 |
| Qwen3-VL-32B | 0.835 | 0.865 | 0.806 | 0.874 |
| *C5:* TEXT(CSV,PARCEL) + PROMPT(MRI) + SWAP IMAGE | | | | |
| InternVL3.5-4B | 0.399 | 0.972 | 0.251 | 0.701 |
| InternVL3.5-14B | 0.703 | 0.902 | 0.576 | 0.808 |
| GLM-4.1V-9B | 0.732 | 0.888 | 0.622 | 0.819 |
| GLM-4.6V-Flash | 0.788 | 0.864 | 0.725 | 0.846 |
| LLaVA-OV-1.5-4B | 0.540 | 0.992 | 0.371 | 0.750 |
| Ministral-3-3B | 0.675 | 0.759 | 0.608 | 0.769 |
| Ministral-3-14B | 0.698 | 0.955 | 0.549 | 0.812 |
| Qwen2.5-VL-3B | 0.728 | 0.666 | 0.803 | 0.763 |
| Qwen2.5-VL-32B | 0.780 | 0.902 | 0.686 | 0.847 |
| Qwen2.5-VL-72B | **0.849** | 0.838 | 0.859 | 0.879 |
| Qwen3-VL-2B | 0.365 | 0.935 | 0.227 | 0.688 |
| Qwen3-VL-32B | 0.822 | 0.893 | 0.762 | 0.870 |

Table 13: Full results on FOR2107 across all five input conditions (C1-C5) and all twelve models. Metrics are F1, Precision (P), Recall (R), and Accuracy (ACC). Best F1 per condition is **bolded**. Left panel: text-only and preamble/image conditions (C1-C3). Right panel: parcellation-augmented and counterfactual swap conditions (C4-C5).

| Model | F1 | P | R | ACC |
|---|---|---|---|---|
| *C1:* TEXT(CSV) | | | | |
| InternVL3.5-4B | 0.616 | 0.470 | 0.893 | 0.594 |
| InternVL3.5-14B | **0.773** | 0.865 | 0.698 | 0.850 |
| GLM-4.1V-9B | 0.278 | 0.909 | 0.164 | 0.689 |
| GLM-4.6V-Flash | 0.359 | 0.901 | 0.224 | 0.708 |
| LLaVA-OV-1.5-4B | 0.680 | 0.591 | 0.799 | 0.725 |
| Ministral-3-3B | 0.504 | 0.692 | 0.396 | 0.716 |
| Ministral-3-14B | 0.602 | 0.878 | 0.458 | 0.779 |
| Qwen2.5-VL-3B | 0.262 | 0.961 | 0.152 | 0.689 |
| Qwen2.5-VL-32B | 0.539 | 0.374 | 0.963 | 0.400 |
| Qwen2.5-VL-72B | 0.786 | 0.927 | 0.682 | 0.865 |
| Qwen3-VL-2B | 0.090 | 0.920 | 0.047 | 0.651 |
| Qwen3-VL-32B | 0.582 | 0.945 | 0.421 | 0.780 |
| *C2:* TEXT(CSV) + PROMPT(MRI) | | | | |
| InternVL3.5-4B | 0.636 | 0.509 | 0.848 | 0.646 |
| InternVL3.5-14B | 0.584 | 0.421 | 0.953 | 0.505 |
| GLM-4.1V-9B | 0.397 | 0.696 | 0.277 | 0.692 |
| GLM-4.6V-Flash | 0.465 | 0.835 | 0.322 | 0.730 |
| LLaVA-OV-1.5-4B | 0.615 | 0.464 | 0.912 | 0.584 |
| Ministral-3-3B | 0.541 | 0.425 | 0.745 | 0.540 |
| Ministral-3-14B | 0.555 | 0.387 | 0.982 | 0.426 |
| Qwen2.5-VL-3B | 0.589 | 0.910 | 0.435 | 0.778 |
| Qwen2.5-VL-32B | 0.534 | 0.365 | 1.000 | 0.365 |
| Qwen2.5-VL-72B | **0.778** | 0.902 | 0.684 | 0.858 |
| Qwen3-VL-2B | 0.028 | 1.000 | 0.014 | 0.641 |
| Qwen3-VL-32B | 0.767 | 0.887 | 0.676 | 0.850 |
| *C3:* TEXT(CSV) + PROMPT(MRI) + MRI PLOT | | | | |
| InternVL3.5-4B | 0.634 | 0.502 | 0.860 | 0.638 |
| InternVL3.5-14B | 0.731 | 0.726 | 0.735 | 0.802 |
| GLM-4.1V-9B | 0.608 | 0.827 | 0.481 | 0.774 |
| GLM-4.6V-Flash | 0.705 | 0.915 | 0.573 | 0.825 |
| LLaVA-OV-1.5-4B | 0.620 | 0.468 | 0.920 | 0.589 |
| Ministral-3-3B | 0.592 | 0.498 | 0.729 | 0.633 |
| Ministral-3-14B | 0.652 | 0.522 | 0.867 | 0.662 |
| Qwen2.5-VL-3B | 0.501 | 0.885 | 0.349 | 0.746 |
| Qwen2.5-VL-32B | 0.688 | 0.625 | 0.766 | 0.747 |
| Qwen2.5-VL-72B | **0.776** | 0.828 | 0.731 | 0.846 |
| Qwen3-VL-2B | 0.008 | 1.000 | 0.004 | 0.636 |
| Qwen3-VL-32B | 0.713 | 0.641 | 0.803 | 0.764 |

| Model | F1 | P | R | ACC |
|---|---|---|---|---|
| *C4:* TEXT(CSV,PARCEL) + PROMPT(MRI) + MRI PLOT | | | | |
| InternVL3.5-4B | 0.656 | 0.563 | 0.784 | 0.699 |
| InternVL3.5-14B | 0.738 | 0.783 | 0.698 | 0.819 |
| GLM-4.1V-9B | 0.455 | 0.867 | 0.308 | 0.730 |
| GLM-4.6V-Flash | 0.659 | 0.836 | 0.544 | 0.795 |
| LLaVA-OV-1.5-4B | 0.681 | 0.570 | 0.844 | 0.711 |
| Ministral-3-3B | 0.563 | 0.521 | 0.612 | 0.653 |
| Ministral-3-14B | 0.675 | 0.670 | 0.680 | 0.761 |
| Qwen2.5-VL-3B | 0.595 | 0.799 | 0.474 | 0.765 |
| Qwen2.5-VL-32B | 0.712 | 0.889 | 0.593 | 0.825 |
| Qwen2.5-VL-72B | **0.775** | 0.878 | 0.694 | 0.853 |
| Qwen3-VL-2B | 0.004 | 1.000 | 0.002 | 0.636 |
| Qwen3-VL-32B | 0.728 | 0.860 | 0.630 | 0.828 |
| *C5:* TEXT(CSV,PARCEL) + PROMPT(MRI) + SWAP IMAGE | | | | |
| InternVL3.5-4B | 0.675 | 0.594 | 0.782 | 0.725 |
| InternVL3.5-14B | 0.683 | 0.658 | 0.711 | 0.759 |
| GLM-4.1V-9B | 0.522 | 0.900 | 0.368 | 0.754 |
| GLM-4.6V-Flash | 0.696 | 0.813 | 0.608 | 0.806 |
| LLaVA-OV-1.5-4B | 0.677 | 0.579 | 0.815 | 0.716 |
| Ministral-3-3B | 0.568 | 0.556 | 0.581 | 0.678 |
| Ministral-3-14B | **0.736** | 0.781 | 0.696 | 0.818 |
| Qwen2.5-VL-3B | 0.619 | 0.833 | 0.493 | 0.779 |
| Qwen2.5-VL-32B | 0.713 | 0.870 | 0.604 | 0.822 |
| Qwen2.5-VL-72B | 0.778 | 0.902 | 0.684 | 0.858 |
| Qwen3-VL-2B | 0.008 | 1.000 | 0.004 | 0.636 |
| Qwen3-VL-32B | 0.667 | 0.860 | 0.544 | 0.801 |

Table 14: Full results on OASIS-3 across all five input conditions (C1-C5) and all twelve models. Metrics are F1, Precision (P), Recall (R), and Accuracy (ACC). Best F1 per condition is **bolded**. Left panel: text-only and preamble/image conditions (C1-C3). Right panel: parcellation-augmented and counterfactual swap conditions (C4-C5). Note that Qwen3-VL-2B produces near-degenerate outputs across all conditions on this dataset, and Qwen2.5-VL-32B under C1-C2 exhibits near-constant positive prediction (ACC $\approx$ 0.40, R $\approx$ 1.00), reflecting calibration failure rather than discriminative ability.

| Hyperparameter | Value |
|---|---|
| Rank ($r$) | 64 |
| Alpha ($\alpha$) | 128 |
| Dropout | 0.05 |
| Bias | none |
| Target modules | `q, k, v, o, gate, up, down_proj` |

Table 15: LoRA adapter configuration.

| Hyperparameter | Value |
|---|---|
| *MPO losses* | |
| Loss types | sigmoid, bco_pair, sft |
| Loss weights | 0.8, 0.2, 1.0 |
| KL penalty ($\beta$) | 0.1 |
| *Optimization* | |
| Learning rate | $5 \times 10^{-5}$ |
| LR scheduler | Cosine |
| Warmup steps | 100 |
| Epochs | 3 |
| Effective batch size | 16 |
| Per-device batch size | 1 |
| Gradient accumulation | 16 |
| *Hardware & precision* | |
| Precision | BF16 + TF32 |
| Gradient checkpointing | ✓ |
| GPU | $1 \times$ H200 |
| Training time | $\approx$12 h |
| Sequence truncation | None (image-safe) |

Table 16: MPO training hyperparameters.

| Name | Description |
|---|---|
| Proband | Test subject |
| Datum_Interview | Date interview as stated on the interview form |
| Datum_Fragebogen | Date of the questionnaire |
| Geburtsjahr | Year of birth |
| Alter | Age |
| Geschlecht | Gender |
| Bildungsjahre | Year of education |
| Bildungsjahre_Vater | Year of education father |
| Bildungsjahre_Mutter | Year of education mother |
| BMI | Body mass index |
| BMI_category | Body mass index category |
| UrbanicityScore | Urbanity score |
| AlterMutterBeiGeburt | Mom age at birth |
| AlterVaterBeiGeburt | Dad age at birth |
| Spezifische_Phobie_Typus | What type of phobia is present |
| Spezifische_Phobie_Typus2 | What type of phobia is present? If more than one, found here |
| Group | Patient grouping or diagnosis |
| Specific_phobia_current | Is the patient currently suffering from a specific phobia? |
| Specific_phobia_lifetime | Is the patient suffering from a specific phobia in his lifetime? |
| Eating_Disorder_current | Is the patient currently suffering from an eating disorder? |
| Eating_Disorder_lifetime | Is the patient suffering from an eating disorder in his lifetime? |
| Alcohol_Use_Disorder_Current | Is the patient currently suffering from an alcohol use disorder? |
| Alcohol_Use_Disorder_Lifetime | Is the patient suffering from an alcohol use disorder in his lifetime? |

Table 17: FOR2107 - Demographics and Clinical Information.

| Name | Description |
|---|---|
| **RS-25: Resilience Scale** | |
| RS251 | If I have plans, I follow them through. |
| RS252 | I usually manage everything somehow. |
| RS253 | I can rely on myself rather than on others. |
| RS254 | It is important for me to stay interested in many things. |
| RS255 | If I have to, I can be alone. |
| RS256 | I am proud of what I have already achieved. |
| RS257 | I'm not easily thrown off track. |
| RS258 | I like myself. |
| RS259 | I can manage several things at the same time. |
| RS2510 | I am determined. |
| RS2511 | I rarely ask myself questions about meaning. |
| RS2512 | I take things as they come. |
| RS2513 | I can get through difficult times because I know I have done it before. |
| RS2514 | I have self-discipline. |
| RS2515 | I stay interested in many things. |
| RS2516 | I often find something to laugh about. |
| RS2517 | My belief in myself helps me even in hard times. |
| RS2518 | I can be relied on in emergencies. |
| RS2519 | I can usually see a situation from several perspectives. |
| RS2520 | I can also overcome myself to do things that I don't really want to do. |
| RS2521 | My life has a purpose. |
| RS2522 | I don't insist on things that I can't change. |
| RS2523 | When I'm in a difficult situation, I usually find a way out. |
| RS2524 | I have enough energy to do everything I have to do. |
| RS2525 | I can accept it if not everyone likes me. |
| **PSS: Perceived Stress Scale** | |
| PSS1sf - PSS3sf | In the last month, how often did you feel upset/unable to control things/nervous? |
| PSS4sf - PSS6sf | In the last month, how often were you able to successfully handle problems/changes? |
| PSS7sf - PSS9sf | In the last month, how often did you feel things were going your way/could not fulfill responsibilities? |
| PSS10sf - PSS12sf | In the last month, how often did you feel on top of things/upset about uncontrolled things? |
| PSS13sf - PSS14sf | In the last month, how often were you able to decide how to spend time/feel difficulties piling up? |

Table 18: FOR2107 - Items for the Resilience Scale (RS-25) and Perceived Stress Scale (PSS).

| Name | Description |
|---|---|
| FSozU1 | I have people who can look after my home (flowers, pets) when I'm not there. |
| FSozU2 | There are people who accept me for who I am. |
| FSozU3 | It is important for my friends/relatives to know my opinion on certain things. |
| FSozU4 | I would like more understanding and care from others. |
| FSozU5 | I have a very trusted person whose help I can always count on. |
| FSozU6 | I can borrow tools and food if necessary. |
| FSozU7 | I have friends/relatives who can listen when I need to talk. |
| FSozU8 | I hardly know anyone I like to go out with. |
| FSozU9 | I have friends/relatives who can give me a hug. |
| FSozU10 | If I am ill, I can ask friends/relatives to do important things (e.g. shopping). |
| FSozU11 | If I'm really depressed, I know who I can go to. |
| FSozU12 | I often feel like an outsider. |
| FSozU13 | There are people who share my joys and sorrows. |
| FSozU14 | With some friends/relatives, I can also be quite relaxed. |
| FSozU15 | I have a trusted person who I feel very comfortable around. |
| FSozU16 | I have enough people who really help me when I get stuck. |
| FSozU17 | There are people who stick by me even when I make mistakes. |
| FSozU18 | I would like more security and closeness. |
| FSozU19 | There are enough people with whom I have a really good relationship. |
| FSozU20 | There is a community of people (circle of friends, clique) that I feel drawn to. |
| FSozU21 | I often get good tips from my circle of friends and acquaintances. |
| FSozU22 | There are people to whom I can show all my feelings without it being embarrassing. |

Table 19: FOR2107 - Questionnaire on Social Support (FSozU) Items.

| Name | Description |
|---|---|
| LEQ_pn1 / LEQ1 | Health: Serious illness of one's own (Type of influence / Influence on life) |
| LEQ_pn2 / LEQ2 | Health: Major change in eating habits |
| LEQ_pn3 / LEQ3 | Health: Major change in sleeping habits |
| LEQ_pn4 / LEQ4 | Health: Significant change in the type or amount of leisure activities |
| LEQ_pn5 / LEQ5 | Health: Major dental procedure |
| LEQ_pn6 / LEQ6 | Health: Pregnancy |
| LEQ_pn7 / LEQ7 | Health: Miscarriage or abortion |
| LEQ_pn8 / LEQ8 | Health: Onset of menopause |
| LEQ_pn9 / LEQ9 | Health: Major difficulties with contraceptive aids |
| LEQ_pn10 / LEQ10 | Work: Difficulties in finding work |
| LEQ_pn11 / LEQ11 | Work: Taking up work outside the home |
| LEQ_pn12 / LEQ12 | Work: Changing to a new type of work |
| LEQ_pn13 / LEQ13 | Work: Changing your working hours or conditions |
| LEQ_pn14 / LEQ14 | Work: Changing your job responsibilities |
| LEQ_pn15 / LEQ15 | Work: Difficulties at work with your employer or other employees |
| LEQ_pn16 / LEQ16 | Work: Major company reorganisations |
| LEQ_pn17 / LEQ17 | Work: Being dismissed or laid off from work |
| LEQ_pn18 / LEQ18 | Work: Ending your working life |
| LEQ_pn19 / LEQ19 | Work: Learning at home or distance learning |
| LEQ_pn20 / LEQ20 | School/Education: Starting or ending a school or training program |
| LEQ_pn21 / LEQ21 | School/Education: Changing schools or training programs |
| LEQ_pn22 / LEQ22 | School/Education: Changing a career goal or major in college |
| LEQ_pn23 / LEQ23 | School/Education: Problems in a school or training program |
| LEQ_pn24 / LEQ24 | Residence: Difficulties in finding accommodation |
| LEQ_pn25 / LEQ25 | Residence: Moving within the same town or city |
| LEQ_pn26 / LEQ26 | Residence: Moving to another town, state, or country |
| LEQ_pn27 / LEQ27 | Residence: Significant changes to your living circumstances |
| LEQ_pn28 / LEQ28 | Love/Partnership: Beginning of a new, close, personal relationship |

Table 20: FOR2107 - Life Experiences Questionnaire (LEQ) Part 1. Note: 'pn' designates Type of Influence, while the number alone designates Influence on Life.

| Name | Description |
|---|---|
| LEQ_pn29 / LEQ29 | Love/Partnership: Entering into an engagement |
| LEQ_pn30 / LEQ30 | Love/Partnership: Problems with boyfriend or girlfriend |
| LEQ_pn31 / LEQ31 | Love/Partnership: Separation from boyfriend/girlfriend or breaking engagement |
| LEQ_pn32 / LEQ32 | Love/Partnership: Pregnancy of wife or girlfriend |
| LEQ_pn33 / LEQ33 | Love/Partnership: Miscarriage or abortion of wife or girlfriend |
| LEQ_pn34 / LEQ34 | Love/Partnership: Marriage or domestic partnership |
| LEQ_pn35 / LEQ35 | Love/Partnership: Change in closeness to partner |
| LEQ_pn36 / LEQ36 | Love/Partnership: Infidelity |
| LEQ_pn37 / LEQ37 | Love/Partnership: Conflict with in-laws |
| LEQ_pn38 / LEQ38 | Love/Partnership: Separation from spouse or partner due to arguments |
| LEQ_pn39 / LEQ39 | Love/Partnership: Separation from spouse/partner due to work, travel, etc. |
| LEQ_pn40 / LEQ40 | Love/Partnership: Reconciliation with spouse or partner |
| LEQ_pn41 / LEQ41 | Love/Partnership: Divorce |
| LEQ_pn42 / LEQ42 | Love/Partnership: Changes in spouse/partner's activities outside the home |
| LEQ_pn43 / LEQ43 | Family/Friends: Addition of a new family member |
| LEQ_pn44 / LEQ44 | Family/Friends: Moving out of a child or family member |
| LEQ_pn45 / LEQ45 | Family/Friends: Major changes in health/behavior of family member or friend |
| LEQ_pn46 / LEQ46 | Family/Friends: Death of a spouse or partner |
| LEQ_pn47 / LEQ47 | Family/Friends: Death of a child |
| LEQ_pn48 / LEQ48 | Family/Friends: Death of a family member or close friend |
| LEQ_pn49 / LEQ49 | Family/Friends: Birth of a grandchild |
| LEQ_pn50 / LEQ50 | Family/Friends: Changes in your parents' marital status |
| LEQ_pn51 / LEQ51 | Parenting: Changes in childcare arrangements |
| LEQ_pn52 / LEQ52 | Parenting: Conflicts with spouse or partner over parenthood |
| LEQ_pn53 / LEQ53 | Parenting: Conflicts with child's grandparents over parenthood |
| LEQ_pn54 / LEQ54 | Parenting: Taking on the responsibilities of being a single parent |
| LEQ_pn55 / LEQ55 | Parenting: Custody disputes with former spouse or partner |

Table 21: FOR2107 - Life Experiences Questionnaire (LEQ) Part 2.

| Name | Description |
|---|---|
| LEQ_pn56 / LEQ56 | Personal/Social: Greater personal achievement |
| LEQ_pn57 / LEQ57 | Personal/Social: Important decision regarding your immediate future |
| LEQ_pn58 / LEQ58 | Personal/Social: Changes in your personal habits (clothing, lifestyle, hobbies) |
| LEQ_pn59 / LEQ59 | Personal/Social: Changes in your religious beliefs |
| LEQ_pn60 / LEQ60 | Personal/Social: Changes in your political views |
| LEQ_pn61 / LEQ61 | Personal/Social: Loss or damage to your personal property |
| LEQ_pn62 / LEQ62 | Personal/Social: Gone on a vacation |
| LEQ_pn63 / LEQ63 | Personal/Social: Taking a trip for non-recreational purposes |
| LEQ_pn64 / LEQ64 | Personal/Social: Changes in family gatherings |
| LEQ_pn65 / LEQ65 | Personal/Social: Changes in your social activities (clubs, events, visits) |
| LEQ_pn66 / LEQ66 | Personal/Social: Beginning of new friendships |
| LEQ_pn67 / LEQ67 | Personal/Social: End of a friendship |
| LEQ_pn68 / LEQ68 | Personal/Social: Acquisition or loss of a pet |
| LEQ_pn69 / LEQ69 | Money: Significant change in your financial situation |
| LEQ_pn70 / LEQ70 | Money: Moderate financial commitment (TV, car, etc.) |
| LEQ_pn71 / LEQ71 | Money: Large financial commitment or mortgage |
| LEQ_pn72 / LEQ72 | Money: Cancellation of a mortgage or loan |
| LEQ_pn73 / LEQ73 | Money: Difficulties with creditworthiness |
| LEQ_pn74 / LEQ74 | Crime/Legal: Victim of theft or identity theft |
| LEQ_pn75 / LEQ75 | Crime/Legal: Victim of a violent crime (rape, assault, etc.) |
| LEQ_pn76 / LEQ76 | Crime/Legal: Involvement in an accident |
| LEQ_pn77 / LEQ77 | Crime/Legal: Involvement in a legal dispute |
| LEQ_pn78 / LEQ78 | Crime/Legal: Involvement in a misdemeanor (tickets, disturbing the peace) |
| LEQ_pn79 / LEQ79 | Crime/Legal: Trouble with the law resulting in arrest or detention |
| LEQ_pn80 / LEQ80 | Other recent experiences having an impact on life (1) |
| LEQ_pn81 / LEQ81 | Other recent experiences having an impact on life (2) |
| LEQ_pn82 / LEQ82 | Other recent experiences having an impact on life (3) |

Table 22: FOR2107 - Life Experiences Questionnaire (LEQ) Part 3.

| Name | Description |
|---|---|
| SozDemo1 – SozDemo5 | Current living, work situation, occupation, social contacts |
| Haushaltu14 / Haushaltab14 | People in the household under/over 14 years |
| Haushaltsnetto | Household net income |
| Schulabschluss | Highest educational qualification achieved by subject |
| Schule_Vater / Schule_Mutter | Highest educational qualification achieved by father / mother |
| GebJahr_Mutter / GebJahr_Vater | Mother's / Father's year of birth |
| Immigration | Own immigration or parents'? |
| Kinder | Do you have children? |
| Soehne_leibl (_age) | Number (and age) of biological sons |
| Toechter_leibl (_age) | Number (and age) of biological daughters |
| Soehne_adopt (_age) | Number (and age) of adopted sons |
| Toechter_adopt (_age) | Number (and age) of adopted daughters |
| Geschwister | Do you have siblings? |
| Brueder_GE (_age) | Number (and age) of brothers (parents shared) |
| Schwestern_GE (_age) | Number (and age) of sisters (parents shared) |
| Halbbrueder / Halbschwestern | Number (and age) of half-brothers / half-sisters |
| Stiefbrueder / Stiefschwestern | Number (and age) of step-brothers / step-sisters |
| Zwillinge_Famil / Zwilling_selbst | Are there twins in your first degree family? / Are you a twin? |
| SS_Stadt / SS_Bundesland | City/State mother lived during pregnancy |
| SS_Risiken1 – SS_Risiken7 | Pregnancy risks (infection, alcohol, drugs, malnutrition, smoking) |
| Geburtskomplikationen1 – 4 | Birth complications (forceps, vacuum, cesarean, other) |
| Geburtsgewicht / SSW_Geburt | Birth weight in grams / Week of birth |

Table 23: FOR2107 - Socio-Demographics, Family Composition, and Pregnancy Variables.

| Name | Description |
|---|---|
| **FzT: Drinking Habits** | |
| FzT1 – FzT5 | Questions on frequency and quantity of alcohol consumption (now/past). |
| FzT6 – FzT9 | Questions on binge drinking (6+ drinks) and inability to stop (now/past). |
| FzT10 – FzT13 | Failing expectations due to alcohol; morning drinking (now/past). |
| FzT14 – FzT17 | Guilt, remorse, and memory loss due to alcohol (now/past). |
| FzT18 – FzT20 | Injuries, advice to reduce, and professional help sought for alcohol. |

Table 24: FOR2107 - Alcohol Use (FzT).

| Name | Description |
|---|---|
| **Subject Demographics & Family** | |
| OASISID | OASIS subject ID |
| GENDER | Subject's gender |
| RACE | Subject's race |
| HAND | Subject's Handedness |
| TWIN | Is this a new informant? (Note: Label suggests 'Twin' but description asks about new informant status) |
| SIBS | How many full siblings does the subject have? |
| KIDS | How many biological children did the subject have? |
| **Living Situation & Independence** | |
| LIVSIT | What is the subject's living situation? |
| LIVSITUA | Living situation (detailed categorization) |
| INDEPEND | What is the subject's level of independence? |
| RESIDENC | What is the subject's primary type of residence? |
| MARISTAT | Subject's current marital status |

Table 25: OASIS-3 - Subject Demographics, Family, and Living Situation.

| Name | Description |
|---|---|
| INSEX | Informant's sex |
| INHISP | Does informant report being of Hispanic/Latino ethnicity? |
| INHISPOR | If yes, what are the informant's reported origins? |
| INRACE | What does informant report as his/her race? |
| INRASEC | What additional race does informant report? |
| INRATER | What additional race, beyond what was indicated above, does informant report? |
| INEDUC | Informant's years of education |
| INRELTO | What is informant's relationship to subject? |
| INLIVWTH | Does the informant live with the subject? |
| INVISITS | If no, approximate frequency of in-person visits |
| INCALLS | If no, approximate frequency of telephone contact |
| INRELY | Is there a question about the informant's reliability? |

Table 26: OASIS-3 - Informant Demographics and Contact Frequency.

| Name | Description |
|---|---|
| **General Medical History** | |
| HYPERTEN | Hypertension |
| HYPERCHO | Hypercholesterolemia |
| DIABETES | Diabetes |
| B12DEF | B12 deficiency |
| THYROID | Thyroid Disease |
| CVAFIB | Atrial fibrillation |
| CVCHF | Congestive heart failure |
| CVANGIO | Angioplasty/endarterectomy/stent |
| CVBYPASS | Cardiac bypass procedure |
| **Vitals** | |
| WEIGHT | Subject Weight (lbs) |
| HEIGHT | Subject Height (inches) |
| BPSYS | Subject Blood Pressure (sitting) (systolic) |
| BPDIAS | Subject Blood Pressure (sitting) (diastolic) |
| HRATE | Subject resting heart rate (pulse) |

Table 27: OASIS-3 - Subject Medical History, Cardiovascular Conditions, and Vitals.

| Name | Description |
|---|---|
| **Lifestyle, Sleep & Trauma** | |
| TOBAC100 | Smoked more than 100 cigarettes in life? |
| SMOKYRS | Total years smoked |
| ALCOHOL | Substance abuse - alcohol (clinically significant over a 12 month period) |
| TRAUMEXT | Traumatic brain injury with extended loss of consciousness ($\geq$ 5 minutes) |
| TRAUMCHR | Traumatic brain injury with chronic deficit or dysfunction |
| APNEA | Sleep apnea |
| INSOMN | Hyposomnia/insomnia |
| **Sensory & Cognition** | |
| VISION | Without corrective lenses, is the subject's vision functionally normal? |
| VISCORR | Does the subject usually wear corrective lenses? |
| VISWCORR | Is the subject's vision functionally normal with corrective lenses? |
| HEARING | Without a hearing aid(s), is the subject's hearing functionally normal? |
| HEARAID | Does the subject usually wear a hearing aid(s)? |
| HEARWAID | Is the subject's hearing functionally normal with a hearing aid(s)? |

Table 28: OASIS-3 - Lifestyle Factors, Sensory Capabilities.