BRAIN ALIGNMENT OF REASONING AND ACTION REPRESENTATIONS FROM VISION-LANGUAGE AND ACTION MODELS DURING NATURALISTIC GAMEPLAY

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how humans and artificial intelligence systems predict and plan by interacting with their environment is a fundamental challenge at the intersection of neuroscience and machine learning. Most brain-encoding studies focus on aligning artificial models with brain activity during language comprehension or passive visual processing, while interactive brain-alignment studies have to date been largely limited to reinforcement-learning (RL) agents and theory-based models. To address this gap, we study brain alignment of representative models from two foundation-model types, namely vision-language models (VLMs) and large-action models (LAMs), using fMRI recordings from participants playing naturalistic Atari-style video games. Specifically, we examine how action-focused and reasoning-focused prompts shape the model's internal representations and their alignment with fMRI brain activity. First, we find that both VLMs and LAMs achieve significantly higher voxel-wise encoding performance than RL baselines, with the advantage holding even under matched feature dimensionality. Second, compared to a "no-prompt" baseline, prompt-driven gains are larger in higher-order frontal-parietal and motor-planning regions than in early visual cortex, roughly 2–2.5$\times$ when averaged over ROI groups, although individual regions are heterogeneous. Third, variance partitioning reveals a qualitatively different representational organization. VLM representations are prompt-symmetric (12.5% unique action vs. 13.6% unique reasoning), whereas LAM is action-dominant (27% unique action vs.$-$5% unique reasoning), with the asymmetry strongest in frontal-motor cortex. Together, these results demonstrate that action-specialized fine-tuning reorganizes multimodal game-state representations in ways that change their alignment with action-associated cortical regions, a difference that is invisible at the level of raw prediction accuracy.

1 INTRODUCTION

Recent work has shown that language-model representations predict human brain activity during language, visual, and multimodal processing, revealing parallels between artificial and neural representations (Toneva & Wehbe, 2019; Schrimpf et al., 2021; Oota et al., 2024b; Conwell et al., 2022; Oota et al., 2025b). Beyond passive tasks, brain activity during interactive gameplay has also been aligned with deep reinforcement-learning (RL) agents, linking model features to neural signals underlying action selection and goal-directed behavior (Tomov et al., 2023; Haberland et al., 2026). However, these studies largely focus on RL or theory-based agents (Van Hasselt et al., 2016; Tsividis et al., 2021), whose representations emphasize action learning and may not capture richer knowledge about objects, relationships, and future outcomes.

This limitation is particularly relevant given growing interest in whether modern vision-language models (VLMs) and large-action models (LAMs) acquire internal world models, structured representations of objects, goals, and environmental dynamics that support prediction and planning (Johnson-Laird, 1983; Ha & Schmidhuber, 2018; Waytowich et al., 2024; Wang et al., 2025b; Xie et al., 2026). Interactive gameplay provides an ideal testbed because it requires perception, rule understanding, multi-step reasoning, and decision-making, allowing us to examine whether these models align with the neural representations supporting human behavior.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

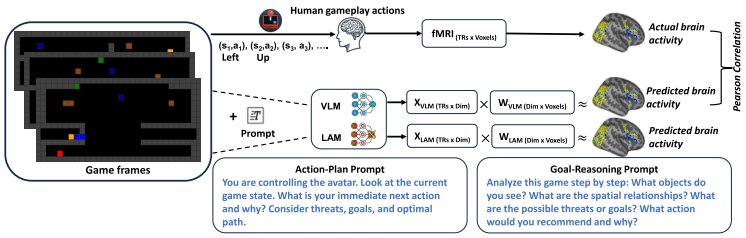


Figure 1: **Brain-alignment pipeline for naturalistic Atari gameplay.** Participants played Atari-style video games during fMRI recording, producing TR-aligned brain responses (top row). The same gameplay frames were processed by two foundation-model families: vision-language models (VLMs) and large-action models (LAMs), conditioning on action-focused or reasoning-focused prompts (bottom). Voxel-wise encoding models trained on per model per prompt representations, X_{VLM} and X_{LAM}, predict held-out fMRI activity, and brain alignment is measured as the Pearson correlation between predicted and actual responses. By comparing brain alignment across model families and prompting conditions, we test how action tuning and task framing shape the alignment of model game-state representations with human brain activity during interactive gameplay.

Successful behavior in humans is thought to rely not only on simple reactive responses, but also on richer internal representations of objects, events, goals, and causal dynamics that enable flexible behavior in new situations. Whether modern AI models acquire analogous representations is therefore an important open question, particularly for VLMs and LAMs, which are increasingly deployed as general-purpose interactive agents and may encode richer perceptual, semantic, and action-oriented structure than traditional RL policies. Aligning these models with human fMRI during gameplay, therefore, offers a principled way to assess not only what information they encode about game states, but also whether their internal representations relate to those that humans engage during interactive play. In this paper, we investigate whether representations from VLMs and LAMs, elicited using action-focused and reasoning-focused prompts, align with human fMRI responses recorded during naturalistic Atari-style gameplay. Hence, we ask the following research questions (RQs): (1) Do VLM and LAM representations align with human brain activity during naturalistic gameplay, and does this alignment differ across action-focused and reasoning-focused prompting conditions? (2) How are the prompt-driven gains (compared to a "no-prompt" setting) in voxel-wise encoding performance distributed across cortex, from early visual to higher-order frontal-parietal regions? (3) How much unique brain-relevant variance is contributed by action- versus reasoning-focused prompts, and how does this decomposition differ between VLMs and LAMs?

To address these questions, we systematically investigate both VLMs and LAMs under reasoning-focused and action-focused prompting conditions, alongside a no-prompt baseline, as shown in Fig. 1, and quantify brain alignment across all three settings. Using brain recordings from 32 participants playing naturalistic Atari-style video games (Tomov et al., 2023), we measure the brain alignment of two VLMs (Qwen-2.5-VL (Qwen Team, 2025) and InternVL3 (Chen et al., 2024)) and two LAMs (UI-TARS-7B-DPO (Qin et al., 2025; Wang et al., 2025a) and OS-Atlas-Pro-7B (Wu et al., 2024)). All models process visual game frames together with text prompts, enabling interactive reasoning over gameplay. We additionally probe Qwen3.5 (Qwen Team, 2026), a thinking-mode model that produces explicit chain-of-thought traces, to probe whether the model's generated reasoning aligns with brain activity differently than its action output. We further compare all models against two RL baselines, EMPA (Tsividis et al., 2021) and D-DQN (Van Hasselt et al., 2016).

Our findings lead to four insights. (1) Both VLMs and LAMs significantly outperform RL baselines (EMPA, D-DQN) in voxel-wise brain encoding, even at matched feature dimensionality; prompting further improves alignment in both. (2) Prompt-driven gains are larger in higher-order frontal-parietal and motor-planning regions (supplementary motor area (SMA), middle frontal gyrus (MFG), inferior frontal gyrus (IFG), angular gyrus (AG)) than in early visual cortex, by roughly 2–2.5$\times$ when averaged over ROI groups, although individual regions are heterogeneous. (3) At matched prompts, whole-brain prediction accuracy does not differ significantly between VLMs and LAMs, but variance partitioning reveals a qualitatively different representational organization. VLMs are prompt-symmetric while LAMs are action-dominant[1], with reasoning becoming redun-

[1] "prompt-symmetric" $\Longrightarrow$ action and reasoning prompts contribute similar amounts. "action-dominant" $\Longrightarrow$ unique contribution of the action prompt substantially exceeds that of the reasoning prompt.

dant in frontal-motor cortex, a dissociation invisible at the level of raw accuracy. (4) In Qwen3.5, representations read out from generated reasoning traces tend to be less brain-aligned than action-output representations, a trend that holds for both last-token and mean-pooled readouts, motivating further investigation of reasoning-trace structure in thinking-mode models.

Together, our results extend brain-alignment research from passive language and visual processing to interactive gameplay, and provide the first systematic brain alignment evaluation of modern VLMs and LAMs in this setting. By contrasting reasoning-focused and action-focused prompts, we operationalize a distinction central to world-model frameworks: the separation between understanding the environment and committing to actions within it. Our findings show that this distinction maps onto qualitatively different representational structures in VLMs and LAMs, dissociations revealed only through variance decomposition. This work opens new directions for studying how foundation models encode the structured, predictive representations that underlie adaptive behavior. Detailed related work on brain alignment, foundation models, and RL models of decision-making is in App. A.

2 DATASET

We use the publicly available raw BIDS-formatted fMRI data from the naturalistic Atari-style video game dataset introduced by Tomov et al. (2023), comprising recordings from 32 healthy, right-handed participants (15 female and 17 male; aged 19–36 years). We preprocess these data using fMRIPrep v25.2.0 (Esteban et al., 2019; 2020); complete preprocessing, nuisance-regression and surface-projection details are in App. B. Each participant completed a single session of $\sim$2.5 hours ($\sim$1.5 hours in scanner and $\sim$1 hour on BOLD acquisition during gameplay), comprising six functional runs, in which participants played six Atari-style video games. Each run contained three blocks, and each block consisted of three levels of increasing difficulty, of a single game; the levels were replayed for a fixed 1-minute duration (nine levels per game total). The runs included a 10-second fixation at the beginning and end, and each run lasted 566 s. For each participant, the fMRI scan was acquired at TR (Repetition Time) = 2 s, producing 283 time points per functional run. In this work, we aggregate across the three levels and focus on cross-game generalization.

For the surface-based ROI analyses, we use the multimodal parcellation of the human cerebral cortex based on the Glasser Atlas (which consists of 180 regions of interest in each hemisphere) to report the ROI (region of interest) analysis for the brain maps (Glasser et al., 2016). We perform the ROI analysis for the Atari video game dataset considering the following language, visual and motor processing regions: angular gyrus (AG), inferior frontal gyrus triangularis (IFGtriang), inferior frontal gyrus opercularis (IFGoperc), middle frontal gyrus (MFG), inferior occipital gyrus (IOG), middle occipital gyrus (MOG), superior occipital gyrus (SOG), calcarine fissure (CAL), lingual gyrus (LING), cuneus (CUN) and supplementary motor area (SMA), based on the prior work (Tomov et al., 2023). We show the flatmap with these labeled ROIs, in Appendix Fig. 6 and list the detailed sub-ROIs of these ROIs in App. C.

3 METHODOLOGY

We use two classes of foundation models for representation extraction: vision-language models (VLMs) and large-action models (LAMs). Both process visual frames together with text prompts.

VLMs. (i) Qwen2.5-VL-7B-Instruct (Qwen Team, 2025) (28 layers, hidden dim 3584) is instruction-tuned on large-scale image–text, video-text data with a dynamic-resolution vision encoder. (ii) InternVL3-8B (Chen et al., 2024) (28 layers, dim 3584) pairs a Qwen2.5-7B language model with a different vision encoder (InternViT) and training corpus.

LAMs. (i) UI-TARS-7B- DPO (Qin et al., 2025; Wang et al., 2025a) is fine-tuned from Qwen2-VL on $\sim$50B tokens of GUI/game-interaction trajectories (SFT+DPO). (ii) OS-Atlas-Pro-7B (Wu et al., 2024) is separately trained from Qwen2-VL on a large GUI grounding corpus. Both LAMs share the Qwen2-VL backbone (28 layers, dim 3584). Combined with Qwen2.5-VL (matched at 28 layers, dim 3584), this enables a within-family VLM-vs-LAM contrast that isolates action-affordance and planning supervision while holding architecture, temporal window, and readout fixed.

Baseline Models. We derive TR-aligned regressors from two Atari-gameplay models. From **EMPA** (Tomov et al., 2023), eight per-frame regressors: four continuous: *surprise* (next-state pre-

diction error), *spriteKL* (belief revision over sprite dynamics), *R_GG* (expected reward under the current goal-graph), *num_effects* (interaction-effect count), and four binary theory-revision flags: (*replan*, *theory_change*, *newEffects*, *interaction_change*). Frame values are averaged within each TR and trimmed by 5 TRs per run edge, yielding a $(T, 8)$ matrix per subject. From **DDQN** (Van Hasselt et al., 2016) (25M-step agent), we extract per-episode action sequences and compute normalized action distributions per run. More details about EMPA features reported in App. D.

VLM and LAM Feature Extraction. To compute model-derived predictors for each fMRI TR, we extract hidden-state representations from both VLMs and LAMs. Let $\mathbf{x}_{t,j}$ denote the j-th frame sampled for fMRI TR t, and let p denote the text prompt. We use $k = 4$ frames per TR. The game frame rate is approximately 20 FPS and the TR is 2 s, corresponding to approximately 40 game frames per TR. For each TR, the window ends at the last game frame in that TR, and four context frames are sampled evenly at intervals of approximately 10 frames (≈ 0.5 s), spanning the final approximately 1.5 s of the TR, denoted as $\mathbf{X}_t = \big(\mathbf{x}_{t,1}, \mathbf{x}_{t,2}, \mathbf{x}_{t,3}, \mathbf{x}_{t,4}\big)$. Frames are taken from each participant's own gameplay, and windows never cross run boundaries. The k sampled frames are jointly passed as a multi-image sequence alongside p, exploiting the model's native multi-image input. We run `model.generate()` with greedy decoding and hidden-state outputs enabled, extracting the last hidden state at every transformer layer to yield one last-token embedding $\mathbf{E}_t \in \mathbb{R}^{(L+1)\times d}$ over all $L + 1$ layers (including the embedding layer).

Episode-level features. For an episode with T TRs, we slide the window across the episode and stack per-TR embeddings into $\mathbf{E} = \{\mathbf{E}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times(L+1)\times d}$. This tensor serves as the predictor matrix for voxel-wise encoding, placing VLM and LAM features on the same temporal footing as EMPA and DDQN. Full notation, chat-template serialization, and per-layer indexing are in App. I.

Prompt variants. To probe different forms of model reasoning, we use two prompt templates. The *action prompt* p_{action} asks the model to commit to a next action and briefly justify the choice (considering threats, goals, and optimal path); the *reasoning prompt* $p_{\text{reasoning}}$ asks the model to analyze the frame step by step (objects, spatial relationships, threats and goals) before recommending an action. Both prompts therefore elicit an action recommendation; they differ in whether the response is organized action-first or reasoning-first. In the *no-prompt* condition, the model receives the same frames with an empty text prompt. Full prompt templates are provided in App. E. For each prompt $p_k \in \{p_{\text{action}}, p_{\text{reasoning}}\}$, we compute $\mathbf{E}_t^{(k)}$ separately and treat the resulting representations as distinct feature sets for downstream voxel-wise encoding.

Prompt controls. In addition to two primary prompt conditions, we evaluate seven control prompts to determine whether the observed effects reflect text presence, visual description, goal information, or motor commitment. These comprise `scrambled-neutral`, which contains task-irrelevant words with syntax destroyed; `scrambled`, which retains the task lexicon but destroys syntax; `lowlevel`, which requests only colour statistics; `scene`, which requests object identities and spatial relationships without goals; `goal`, which requests hazards and objectives without a motor commitment; and `action-bare`, which requests only the selected movement key; and `action-full`, which requests a committed action justified against alternatives. Together with the two primary prompts, these form the nine-condition prompt battery reported in Appendix Table 1.

4 Experimental Setup

Voxel-wise encoding model. We train voxel-wise fMRI encoding models using bootstrap ridge regression (Tikhonov & Arsenin, 1977) to predict voxel responses from model representations. Feature channels and fMRI responses are z-scored separately for training and testing. To account for hemodynamic delay, we concatenate four temporal lags (2, 4, 6, and 8 s), allowing voxel-specific lag weights to be learned (Toneva & Wehbe, 2019). At time step t, let stimulus features be $X_t \in \mathbb{R}^{N\times D}$ and voxel responses $Y_t \in \mathbb{R}^{N\times V}$, where N is the number of examples, D the lagged feature dimension, and V the number of voxels. Regularization parameters are selected via cross-validation. The goal is to predict each voxel's brain response from the presented stimulus.

Layer selection. We fit encoding models for every layer and report the best-performing layer per model. Because each layer's correlation is itself cross-validated but the layer is chosen using these same held-out scores, the reported peak values are slightly optimistic. Hence, we report full layer-wise profiles (App. Figs. 11–16) and verify that our conclusions don't depend on selected layer.

Train-test setup. We train subject-specific encoding models using strict leave-one-run-out (LORO) cross-validation. For each subject, in each fold, one complete scanner run is held out as the test set, while all remaining runs are used for training. This process is repeated until every run has served once as the held-out test set. Feature windows are constructed separately within each run and never cross run boundaries. Thus, no test TR shares a scanner run with the training data, preventing within-run train-test adjacency and reducing potential information leakage due to temporal autocorrelation. The detailed hyperparameter settings and statistical significance tests are provided in App. F and H.

Evaluation metrics. We evaluate our models using Pearson Correlation Coefficient (PCC), which is a standard metric to evaluate brain alignment (Jain & Huth, 2018; Schrimpf et al., 2021; Goldstein et al., 2022). Let TR be the number of time repetitions in the test set. Let $Y = \{Y_i\}_{i=1}^{TR}$ and $\hat{Y} = \{\hat{Y}_i\}_{i=1}^{TR}$ denote the actual and predicted value vectors for a single voxel, respectively. Thus, Y and $\hat{Y} \in \mathbb{R}^{TR}$. PCC is then calculated as correlation between the model's predictions $\hat{Y}$ and neural recordings Y. Our primary brain-alignment analysis retains all cortical vertices without applying a predictivity threshold. We additionally report descriptive sensitivity analyses, and report averages over voxels with $r > 0$ and $r > 0.05$. We report mean $\pm$ SEM across the 32 participants. Model and prompt comparisons use paired tests across participants, with FDR correction across ROIs; details are in App. H. Implementation details are in App. F.

Variance partitioning. Using a variance partitioning approach (de Heer et al., 2017; LeBel et al., 2021), we test whether action- and reasoning-based prompts carry overlapping or distinct brain-relevant information, and whether LAMs, which are fine-tuned on action-specific environments reshapes this decomposition. Implementation details are in App. N.

Behavioral analyses. To relate model representations to participants' behavior, we (i) compare the action recommended in each model's generated output with the participant's action at the same time point (model–human action agreement, evaluated at 2 Hz), and (ii) decode each participant's TR-level actions from the same model features used for encoding, using a linear classifier under leave-one-run-out cross-validation. Details are in App. J.

5 RESULTS

[RQ1]: VLMS AND LAMS EXCEED RL BASELINES UNDER *no-prompt* CONDITIONS, AND TASK-FRAMED PROMPTS FURTHER IMPROVE BRAIN ALIGNMENT.

To test whether VLMs and LAMs achieve higher voxel-wise encoding performance than classical RL baselines, and to isolate this from confounds of prompting and feature dimensionality, we evaluate two settings. First, a *no-prompt* setting in which VLMs and LAMs receive only gameplay frames without text instruction; this isolates representational quality from prompt-induced effects and offers fair comparison with EMPA and DDQN, neither of which uses language input. Within this setting, we additionally vary feature dimensionality (8, 64, 1024) to rule out that any encoding gain simply reflects VLMs and LAMs having more features than the RL baselines. Second, a *prompted* setting in which the same models additionally receive text-prompt instructions (full templates in App. E), quantifying the gain from explicit prompting. Across both settings, VLM and LAM representations significantly outperform RL baselines, with prompting providing additional gains. We additionally test whether the prompted-alignment results generalize across two VLMs and two LAMs, quantify model–human action agreement, and examine whether participants' actions can be decoded from the model representations.

VLMs and LAMs outperform RL baselines under "*no-prompt*" conditions. As shown in Fig. 2 (left), both VLM and LAM representations outperform the classical baselines in whole-brain voxel-wise encoding without any prompt instruction (shown for all voxels unthresholded). Specifically, VLM (*no-prompt*) achieves r=0.029 and LAM (*no-prompt*) reaches r=0.027, compared to EMPA (r=0.013) and DDQN (r=0.005). The advantage is consistent across all three feature dimensionalities tested. Crucially, even when VLM and LAM features are reduced to 8 dimensions to match EMPA's feature count, they continue to significantly show higher brain alignment than both RL baselines (VLM r=0.021, LAM r=0.020), ruling out feature count as the source of the gain. Performance saturates between 64-dim and 1024-dim, suggesting that the gain over RL baselines reflects what the features encode rather than feature dimensionality alone. Together, these results indicate that contemporary multimodal architectures learn richer latent representations of dynamic game en-

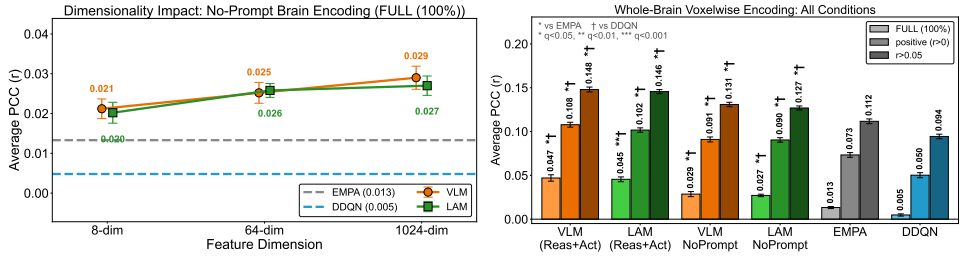


Figure 2: **Whole-brain voxel-wise encoding performance.** (left) Effect of feature dimensionality under the *no-prompt* condition, shown for voxels with full voxels (unthresholded): VLM (orange) and LAM (green) features evaluated at three dimensionalities (8, 64, 1024), compared against the RL baselines EMPA (gray dashed, r=0.013) and DDQN (light blue dashed, r=0.005). * indicates significant pairwise gain from 8 to 64 dim (paired t-test, $p < 0.05$); the 64 to 1024 transition is non-significant, indicating saturation. At every dimensionality tested, both VLM and LAM significantly exceed both RL baselines. (right) All conditions: Whole-brain alignment for prompted (average across Reasoning and Action) and *no-prompt* VLM and LAM, compared to EMPA and DDQN baselines, under three voxel-selection criteria: all voxels (Full), positively predicted (r>0) and (r>0.05). * denotes a significant difference vs. EMPA and † vs. DDQN (paired t-test, FDR-corrected; * $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$). Error bars: mean ± SEM across subjects.

vironments than traditional task-optimized RL systems. We present similar analysis for voxels with $r \geq 0.05$ in App. M Fig. 9.

Effect of Action and Reasoning prompts. We next examine whether explicit prompting further improves brain alignment by eliciting higher-level reasoning states from the same models. Fig. 2 (right) reports all conditions on the primary unthresholded metric alongside the two sensitivity criteria. Under action and reasoning prompts, as shown in Fig. 2 (right), VLM and LAM achieve significantly higher encoding performance than their *no-prompt* counterparts (paired t-test, p<0.05): VLM reaches r=0.148 and LAM reaches r=0.146. At the unthresholded whole-cortex level, prompted VLM features predict brain activity significantly better than EMPA (VLM $r = 0.047 \pm 0.027$ vs. EMPA $r = 0.013 \pm 0.012$, mean ± SEM across $N = 32$ participants; $\Delta r = 0.034$; paired $t(31)$, $p = 0.047$; Wilcoxon signed-rank $p = 0.033$), and the LAM shows the same advantage ($r = 0.045$; $q < 0.01$ vs. EMPA). This shows that adding task-framed text instructions improves brain alignment beyond what gameplay frames alone provide. Prompted VLM and LAM achieve comparable whole-brain alignment, with the difference between them not reaching significance. We further investigate representational differences masked by encoding accuracy using variance partitioning in Section 5. A second VLM–LAM pair (InternVL3, OS-Atlas) yields comparable alignment under both prompts, with small, non-significant prompt preferences (≤ 0.006; App. K Fig. 7); we focus subsequent analyses on Qwen2.5-VL and UI-TARS.

Prompt controls. To test whether prompt effects reflect text presence, visual description, goal information, or motor commitment alone, we evaluate seven control prompts (Table 1). Scrambled text, low-level colour descriptions, and scene descriptions yield weaker alignment than the task-framed prompts ($r \approx 0.006$–0.029), whereas action-full and goal prompts approach them. Effects are model-dependent (e.g., action-bare: LAM ≈ 0.055, VLM ≈ 0.029), so we do not interpret individual prompts as isolating distinct semantic dimensions (App. Fig. 8).

Behavior analysis. Model features decode participants' TR-level actions above chance (balanced accuracy: VLM 0.342, LAM 0.362, EMPA 0.289). Details in App. J.

[RQ2]: PROMPT-DRIVEN GAINS ARE SIGNIFICANTLY GREATER IN HIGHER-ORDER THAN IN EARLY VISUAL CORTEX, AND HIGHER-ORDER ASSOCIATION REGIONS TEND TO ALIGN BEST WITH DEEPER MODEL LAYERS.

For the primary unthresholded analysis, we group the 11 ROIs into two functional groups: (1) a *higher-order group* spanning frontal-parietal and motor-planning cortex (AG, IFGtriang, IFGoperc, MFG, SMA), and (2) *early-visual group* (CAL, CUN, LING), and a higher-visual group (SOG, IOG, MOG). We examine the higher-visual ROIs as an intermediate group in the descriptive analysis.

Prompt gains differ between early-visual and higher-order ROI groups. Fig. 3 reports unthresholded ROI means for the VLM under reasoning, action, and *no-prompt* conditions, together with the

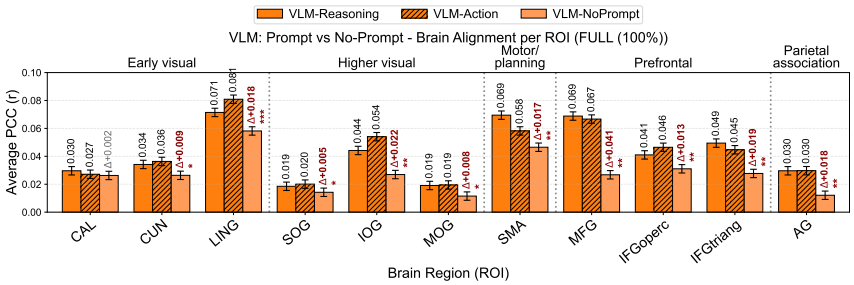


Figure 3: Brain alignment averaged across participants and voxels per ROI for prompted vs. no-prompt VLM representations. Average Pearson correlation between predicted and observed fMRI responses across 11 ROIs for VLM-Reasoning (solid), VLM-Action (hatched), and VLM-NoPrompt (light). Error bars denote mean ± SEM across subjects. Markers denote paired comparisons against no-prompt (R: Reasoning, A: Action; *p $<$ 0.05, **p $<$ 0.01). Both prompted conditions show higher alignment than no-prompt across all ROIs. Similar plot for LAM in Appendix Fig. 14.

composite prompt gain $\Delta r = r_{\text{prompted}} - r_{\text{no-prompt}}$, where r_{prompted} is averaged across the reasoning and action conditions; the corresponding LAM results are shown in App. M Fig. 14. For the VLM, prompting significantly raises alignment in 10 of 11 ROIs (paired t-tests, FDR-corrected). The largest gains occur in MFG ($\Delta r = +0.041$, $q < 0.01$), followed by IOG ($+0.022$, $q < 0.01$), IFGtriang ($+0.019$), LING and AG ($+0.018$), SMA ($+0.017$), IFGoperc ($+0.013$), with smaller gains in CUN ($+0.009$), MOG ($+0.008$), and SOG ($+0.005$) (all $q < 0.05$). CAL ($+0.002$, not significant (ns)). For the LAM, prompting significantly raises alignment in all 11 ROIs. Gains are largest in SMA ($+0.057$, $q < 0.001$) and MFG ($+0.042$, $q < 0.001$), intermediate in IOG and IFGtriang (both $+0.030$), SOG and AG (both $+0.023$), LING ($+0.021$), and MOG ($+0.019$), and smallest in IFGoperc ($+0.016$), CUN ($+0.010$), and CAL ($+0.009$) (all $q < 0.05$). The LAM *no-prompt* baseline is close to zero in SOG, MOG, and AG, so prompted alignment in these regions arises almost entirely from prompting. For both families, prompted alignment is highest in LING, SMA, and MFG and lowest in SOG and MOG.

Early-visual vs. higher-order cortex. To test the group-level regional organization of prompt effects, we grouped ROIs into an early-visual, a higher-visual, and a higher-order group. We computed each participant's composite prompt gain within each ROI group and compared the subject-wise values. Using all voxels, prompt gains are significantly larger in higher-order than in early-visual cortex for both model families (VLM: higher-order minus early-visual $\Delta r = +0.012$, $t(31) = 3.83$, $q = 0.019$; LAM: $\Delta r = +0.020$, $t(31) = 3.31$, $q = 0.039$). Higher-order gains are thus roughly 2–2.5× early-visual gains in both families. In both families, the higher-visual group is intermediate (VLM: $+0.012$; LAM: $+0.024$), yielding an early-visual $<$ higher-visual $<$ higher-order ordering at the ROI-group level. The contrast remains significant when restricting the analysis to vertices with $r > 0.05$ (VLM: $+0.029$ vs. $+0.012$, $\Delta r = +0.018$, $t(31) = 2.85$, $q = 0.041$; LAM: $+0.041$ vs. $+0.016$, $\Delta r = +0.025$, $t(31) = 2.13$, $q = 0.046$), where the higher-visual group is likewise intermediate (VLM: $+0.013$; LAM: $+0.028$). These higher-order ROIs partially overlap with prefrontal regions whose activity Tomov et al. (2023) found to be predicted by EMPA's theory-based representations, providing anatomical correspondence with prior findings from the same dataset.

These results support a modest macro-scale regional effect rather than a strictly monotonic cortical hierarchy. Individual ROIs remain heterogeneous: IOG, a higher-visual region, shows gains exceeding several frontal ROIs in both families (VLM: $+0.022$; LAM: $+0.030$), whereas CAL shows the smallest gain in both families. In contrast to prompt gains, absolute prompted alignment differs only slightly between the early-visual and higher-order groups (VLM: $r = 0.047$ vs. 0.050; LAM: $r = 0.045$ vs. 0.051). We therefore restrict our regional conclusion to the significant group-level difference in prompt-induced gains and make no claim about an absolute regional ordering or, beyond the descriptive ROI-group ordering, a monotonic progression across individual ROIs.

Layer-wise and model comparisons. Visual ROI groups peak at early layers (VLM L4; LAM L7), whereas motor/planning and prefrontal groups peak at intermediate-to-deep layers (L18–L21); these peaks are descriptive (App. Table 7, Figs. 11–16). Raw encoding accuracy does not differ significantly between VLM and LAM in any ROI after FDR correction.

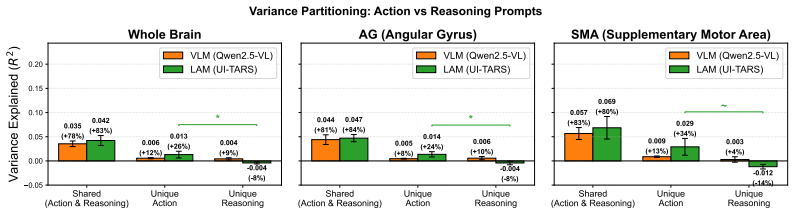


Figure 4: **Variance partitioning of action vs. reasoning prompts.** R^2 decomposed into shared variance and prompt-unique components for VLM (Qwen2.5-VL, orange) and LAM (UI-TARS-7B-DPO, green), averaged across participants. Numbers above bars show absolute R^2 and (in parentheses) percentage of joint explained variance. VLM unique variance is balanced across prompts; LAM is action-dominant, with negative unique-reasoning variance in whole brain and SMA. Brackets: VLM-vs-LAM comparison on unique reasoning, * $p < 0.05$, $\sim$ marginal ($p < 0.10$). Error bars: mean $\pm$ SEM across subjects.

VLM–LAM regional comparison. Raw encoding accuracy does not differ significantly between the VLM and LAM in any ROI after FDR correction. Under the action prompt, all ROI-wise comparisons yield $q > 0.90$, while under the reasoning prompt the smallest adjusted value is $q = 0.08$. Within the LAM, action-prompt alignment is numerically equal to or higher than reasoning-prompt alignment in 10 of 11 ROIs, with IFGtriang as the only exception, whereas the VLM exhibits a more mixed regional pattern. However, none of the within-model action-versus-reasoning differences reaches ROI-level significance, so this pattern is descriptive. In RQ3 (Sec. 5), we test whether this numerical action-leaning pattern corresponds to differences in the shared and prompt-unique information encoded by the two model families.

Spatial visualization. Group-averaged across subjects per-voxel prompt gains, r_{prompted}-$r_{\text{no-prompt}}$, under both reasoning (App. M Fig. 15a) and action (App. M Fig. 15b) prompts, reveal a consistent spatial pattern: large positive gains (red) concentrated in frontal-parietal and motor-planning cortex (MFG, SMA, IFG, AG), substantially smaller gains in mid-level visual regions, and a mixture of positive and negative gains within early visual cortex (CAL) and mid-level visual (SOG). The spatial pattern recapitulates the ROI-level ordering and clarifies that the small ROI-mean gains in early visual regions reflect heterogeneous voxel-level effects rather than uniform improvement, consistent with these regions being driven primarily by direct visual-input encoding rather than prompt-conditioned representations.

[RQ3]: VARIANCE PARTITIONING REVEALS PROMPT-SYMMETRIC ALIGNMENT IN VLMS BUT ACTION-DOMINANT BRAIN ALIGNMENT IN LAMS.

Although VLM and LAM achieve comparable raw prediction accuracy at matched prompts (refer [RQ2]), this similarity may mask differences in what the two models encode. To probe this, we apply variance partitioning to decompose the brain variance explained by Reasoning (R) and Action (A) prompts into shared and prompt-unique components, separately for each model. For this, we use standard variance partitioning approach discussed in Section 4. Fig. 4 presents the shared variance between action and reasoning prompts for the VLM and LAM models, averaged across all subjects.

As shown in Fig. 4, for VLM, the two prompts contributed statistically indistinguishable unique variance (paired t-test p = 0.75). For action prompts, u_A = 0.006 (12.4% of explained variance). For reasoning prompts, u_R = 0.004 (9.5% of explained variance). This is over a large shared component (0.035, 78.1% of explained variance), indicating prompt-symmetric representations. In contrast, the LAM exhibits a significant asymmetry. Action prompts contributed substantially more unique variance than Reasoning prompts (u_A = 0.013 (25.6%) vs u_R = -0.004 (-8.2%); p = 0.032), while the shared component remained high (0.042, 82.4%). The LAM's unique-reasoning variance is significantly lower than the VLM's at the whole-brain level ($p < 0.05$). This dissociation is regionally specific. In AG, VLM shows balanced prompt effects (8% action vs. 10% reasoning), whereas LAM shifts toward action-dominant representations, with u_A = 24% of variance and $u_R = -8\%$. In SMA, the contrast is strongest: both VLM unique components are small (13% action vs. 4% reasoning), whereas the LAM devotes one-third of explainable variance to action (u_A = 34%, with u_R=-14%).

Together, these results reveal a clear dissociation. VLM representations are prompt-symmetric (12.4% vs. 9.5% unique variance), whereas LAM representations are action-dominant (25.6% vs.

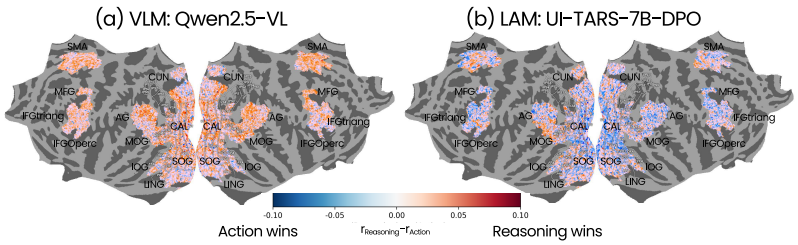


Figure 5: **Spatial visualization confirms the VLM/LAM dissociation.** Group-averaged across subjects per-voxel difference maps ($r_{Reasoning}$ - r_{Action})

-8.2% unique variance). This suggests that action-specialized fine-tuning reorganizes model representations toward action-associated cortical regions, despite similar overall predictive accuracy. Variance-partitioning results for all seven ROIs are reported in App. N, Fig. 19.

Spatial visualization confirms the VLM/LAM dissociation. Group-averaged voxel-wise difference flatmaps across subjects ($r_{\text{Reasoning}} - r_{\text{Action}}$; Fig. 5) show VLMs are reasoning-dominant (red), with positive differences across dorsal stream and lateral occipital cortex and only localized Action-favoring regions in early visual cortex (Fig. 5a). In contrast, LAMs exhibit widespread action-dominant (blue) voxels across early visual and ventral regions, with only scattered Reasoning-favoring voxels dorsally (Fig. 5b). These maps mirror the ROI-level prompt comparisons (Fig. 3) and variance-partitioning (Fig. 4) results, indicating prompt-symmetric VLM representations but action-leaning LAM representations despite comparable prediction accuracy.

Generalization across VLM–LAM pairs. To test whether the prompt-symmetric VLM vs action-dominant LAM pattern generalizes beyond the Qwen-VL family, we additionally ran variance partitioning on second pair InternVL3 (Qwen2.5-VL-based VLM) vs OS-Atlas-Pro (Qwen2-VL-based LAM). The qualitative pattern is preserved. At whole-brain, OS-Atlas shows substantially more u_A variance than u_R (0.0102 vs. 0.002), while InternVL3 is approximately prompt-symmetric (0.004 vs. 0.006). The shared component remained high in LAM vs. VLM (0.031, 0.037). ROI analyses (AG, SMA) show the same OS-Atlas-Pro action-dominance, with InternVL3 shifting slightly toward Reasoning-leaning representations in higher-order cortex. Detailed analysis is in App. N.1.

Thinking-mode readouts. In Qwen3.5, the last-token reasoning readout shows weaker brain alignment than the final-answer readout ($r = 0.012$ vs. 0.031; paired t-test $p = 0.032$), but mean-pooling the reasoning trace raises alignment to $r = 0.036$, indicating that readout position, not trace content, drives most of the difference (App. O).

6 Discussion and Conclusion

We evaluate how VLMs and LAMs align with human brain activity during Atari-style gameplay using voxel-wise encoding on an fMRI dataset. Comparing two VLMs and two LAMs against RL baselines yields four key findings. **First**, both VLMs and LAMs substantially outperform RL baselines, even after matching feature counts, showing that multimodal foundation models encode more brain-relevant information than traditional task-optimized RL agents. **Second**, prompt-driven gains are strongest in higher-order frontal-parietal and motor-planning regions such as MFG, SMA, IFG, and AG, with smaller gains in early visual cortex. This suggests prompts improve representations related to goals, planning, and decision-making, likely reflecting multiple-demand network activity rather than language alone (Duncan, 2010; Fedorenko et al., 2024). **Third**, although VLMs and LAMs achieve similar whole-brain prediction accuracy, variance partitioning reveals distinct internal organization. VLMs show balanced contributions from Action and Reasoning prompts, whereas action-tuned LAMs are strongly action-dominant, especially in frontal-motor regions. This indicates that action fine-tuning already captures much of the information reasoning prompts provide. **Fourth**, in a preliminary extension, Qwen3.5 reasoning-trace representations tend to align less with whole-cortex brain activity than its final-answer representations ($r = 0.012$ vs. 0.031 for last-token readouts), suggesting that explicit chain-of-thought reasoning does not necessarily improve brain alignment. Together, these findings extend brain-alignment research from passive perception to interactive gameplay and uncover representational differences between VLMs and action-tuned models that emerge only through variance decomposition. We discuss limitations in App. P.

AI USE STATEMENT

We did not use generative AI tools for any of the tasks requiring mandatory disclosure under the ICLR AI policy. We acknowledge the use of AI tools (Grammarly, OpenAI GPT-5.5, Claude Opus 4.8) for grammar correction, refining phrasing for better readability, and refining the presentation of figures. All AI-assisted edits were verified and finalized by the authors.

ETHICS STATEMENT

We did not create any new neural recordings data as part of this work. We used the Atari-style video game datas which is publicly available without any restrictions. Atari-style video game datast can be downloaded from `https://openneuro.org/datasets/ds004323/versions/1.0.1`. This dataset is raw BIDS-formatted dataset and we performed preprocessing before conducting experiments. Please read their terms of use[2] for more details.

We do not foresee any harmful uses of this technology.

REPRODUCIBILITY STATEMENT

We provide clear details on the fMRI dataset, the fMRI preprocessing details, and the training procedure in the Methodology. We plan to release the code upon publication.

REFERENCES

Richard Antonello, Javier S Turek, Vy Vo, and Alexander Huth. Low-dimensional structure in the space of language representations is reflected in brain responses. *Advances in Neural Information Processing Systems*, 34:8332–8344, 2021.

Cordell M Baker, Joshua D Burks, Robert G Briggs, Andrew K Conner, Chad A Glenn, Kathleen N Taylor, Goksel Sali, Tressie M McCoy, James D Battiste, Daniel L O'Donoghue, et al. A connectomic atlas of the human cerebrum—chapter 7: the lateral parietal lobe. *Operative Neurosurgery*, 15(suppl_1):S295–S349, 2018.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, 2022.

Catherine Chen, Tom Dupré la Tour, Jack Gallant, Dan Klein, and Fatma Deniz. The cortical representation of language timescales is shared between reading and listening. *bioRxiv*, pp. 2023–01, 2023.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.

Nilearn contributors. nilearn, 2026. URL `https://github.com/nilearn/nilearn`.

Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*, pp. 2022–03, 2022.

Wendy A de Heer, Alexander G Huth, Thomas L Griffiths, Jack L Gallant, and Frédéric E Theunissen. The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, 37(27):6539–6557, 2017.

Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, 2019.

[2] `https://openneuro.org/datasets/ds004323/versions/1.0.1`

Rutvik H Desai, Usha Tadimeti, and Nicholas Riccardi. Proper and common names in the semantic system. *Brain Structure and Function*, 228(1):239–254, 2023.

Dota Tianai Dong and Mariya Toneva. Interpreting multimodal video transformers using brain recordings. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023a.

Dota Tianai Dong and Mariya Toneva. Vision-language integration in multimodal video transformers (partially) aligns with the brain. *arXiv preprint arXiv:2311.07766*, 2023b.

John Duncan. The multiple-demand (md) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences*, 14(4):172–179, 2010.

Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

Oscar Esteban, Rastko Ciric, Karolina Finc, Ross W Blair, Christopher J Markiewicz, Craig A Moodie, James D Kent, Mathias Goncalves, Elizabeth DuPre, Daniel EP Gomez, et al. Analysis of task-based functional mri data preprocessed with fmriprep. *Nature protocols*, 15(7):2186–2202, 2020.

Evelina Fedorenko, Anna A Ivanova, and Tamar I Regev. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25(5):289–312, 2024.

Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

Sabine Haberland, Hannes Ruge, and Holger Frimmel. Encoding neural representations of time-continuous stimulus-response transformations in the human brain with advanced deep neural networks. *Imaging Neuroscience*, 2026.

Shailee Jain and Alexander G Huth. Incorporating context into language encoding models for fmri. In *NIPS*, pp. 6629–6638, 2018.

Philip Nicholas Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, 1983.

Amanda LeBel, Shailee Jain, and Alexander G Huth. Voxelwise encoding models show that cerebellar language representations are highly conceptual. *Journal of Neuroscience*, 41(50):10341–10355, 2021.

Juliette Millet, Charlotte Caucheteux, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, Jean-Remi King, et al. Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, 35: 33428–33443, 2022.

Camille K Milton, Vukshitha Dhanaraj, Isabella M Young, Hugh M Taylor, Peter J Nicholas, Robert G Briggs, Michael Y Bai, Rannulu D Fonseka, Jorge Hormovas, Yueh-Hsin Lin, et al. Parcellation-based anatomic model of the semantic network. *Brain and Behavior*, 11(4):e02065, 2021.

Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. Unveiling multi-level and multi-modal semantic representations in the human brain using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20313–20338, 2024.

Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Surampudi. Neural language taskonomy: Which nlp tasks are the most predictive of fmri brain activity? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3220–3237, 2022a.

Subba Reddy Oota, Jashn Arora, Vijay Rowtula, Manish Gupta, and Raju S Bapi. Visio-linguistic brain encoding. In *COLING*, pp. 116–133, 2022b.

Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. Speech language models lack important brain-relevant semantics. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8503–8528. Association for Computational Linguistics, 2024a. URL https://aclanthology.org/2024.acl-long.462.

Subba Reddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 36, 2024b.

Subba Reddy Oota, Akshett Rai Jindal, Ishani Mondal, Khushbu Pahwa, Satya Sai Srinath Namburi GNVV, Manish Shrivastava, Maneesh Kumar Singh, Bapi Raju Surampudi, and Manish Gupta. Correlating instruction-tuning (in multimodal models) with vision-language processing (in the brain). In *The Thirteenth International Conference on Learning Representations*, 2025a.

Subba Reddy Oota, Khushbu Pahwa, mounika marreddy, Maneesh Kumar Singh, Manish Gupta, and Bapi Raju Surampudi. Multi-modal brain encoding models for multi-modal stimuli. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025.

Team Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Team Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

A. D. Ritchie. The nature of explanation. by k. j. w. craik. (cambridge university press. 1943. pp. viii + 123. price 6s.). *Philosophy*, 19(73):173–174, 1944. doi: 10.1017/S0031819100004733.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 2021.

Andreĭ Nikolaevich Tikhonov and VIAK Arsenin. Solutions of ill-posed problems. *(No Title)*, 1977.

Momchil S Tomov, Pedro A Tsividis, Thomas Pouncy, Joshua B Tenenbaum, and Samuel J Gershman. The neural architecture of theory-based reinforcement learning. *Neuron*, 111(8):1331–1344, 2023.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, 32, 2019.

Pedro A Tsividis, Joao Loula, Jake Burga, Nathan Foss, Andres Campero, Thomas Pouncy, Samuel J Gershman, and Joshua B Tenenbaum. Human-level reinforcement learning through theory-based modeling, exploration, and planning. *arXiv preprint arXiv:2107.12544*, 2021.

Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *Plos Biology*, 21(12):e3002366, 2023.

Aditya Vaidya, Shailee Jain, and Alexander Huth. Self-supervised models of audio effectively explain human cortical responses to speech. In *International Conference on Machine Learning*, pp. 21927–21944. PMLR, 2022.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Haoming Wang, Haoyang Zou, Huatong Song, Jiazhan Feng, Junjie Fang, Junting Lu, Longxiang Liu, Qinyu Luo, Shihao Liang, Shijue Huang, et al. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning. *arXiv preprint arXiv:2509.02544*, 2025a.

Xinyu Wang, Bohan Zhuang, and Qi Wu. Are large vision language models good game players? In *The Thirteenth International Conference on Learning Representations*, 2025b.

Nicholas R Waytowich, Devin White, MD Sunbeam, and Vinicius G Goecks. Atari-gpt: Benchmarking multimodal large language models as low-level policies in atari games. *arXiv preprint arXiv:2408.15950*, 2024.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. Os-atlas: A foundation action model for generalist gui agents. *arXiv preprint arXiv:2410.23218*, 2024.

Yunfei Xie, Yinsong Ma, Shiyi Lan, Alan Yuille, Junfei Xiao, and Chen Wei. Play to generalize: Learning to reason through game play. *ICLR*, pp. arXiv–2506, 2026.

Overview of Appendix Sections

- App. A: Related Work
- App. B: fMRI Preprocessing
- App. C: Detailed sub-ROIs of language, visual and auditory regions
- App. D: Baseline Models Features
- App. E: Prompt Templates for Brain Encoding Models
- App. F: Hyperparameter Details
- App. G: Details of explained variance partitioning
- App. H: Statistical significance
- App. I: Detailed Feature Extraction
- App. J: Behavioral Analysis
- App. K: Generalization across model families
- App. L: Prompt Controls
- App. M: VLM and LAM: Prompt-driven alignment gains across ROIs
- App. N: Variance Partitioning Results
- App. O: Alignment using generated reasoning traces of thinking-mode models.
- App. P: Limitations

A Related Work

Foundation models and brain alignment. Our work also relates to a growing literature that investigates the alignment between human brains and language models. A number of studies have used text-based language models have been shown to predict both text- and speech-evoked brain responses with high fidelity (Jain & Huth, 2018; Deniz et al., 2019; Toneva & Wehbe, 2019; Caucheteux & King, 2022; Antonello et al., 2021; Oota et al., 2022a; 2024b). Advances

in Transformer-based speech models have similarly motivated work on speech-evoked brain alignment (Millet et al., 2022; Vaidya et al., 2022; Tuckute et al., 2023; Oota et al., 2024a; Chen et al., 2023). More recently, multimodal Transformers have been tested under both unimodal (Dong & Toneva, 2023a; Oota et al., 2022b; 2025a) and multimodal naturalistic stimuli (Dong & Toneva, 2023b; Nakagi et al., 2024; Oota et al., 2025b). Our work is complementary: we apply recent foundation models, specifically vision-language and large-action models, to study brain alignment during interactive Atari-style gameplay, with a particular focus on how action and reasoning prompt-conditioned representations differ across the cortical hierarchy.

Reinforcement learning, world models, and brain alignment during interactive behavior. A complementary line of work has aligned brain activity recorded during interactive gameplay with deep reinforcement-learning (RL) agents and theory-based models (Tomov et al., 2023; Haberland et al., 2026). Notably, Tomov et al. (2023) introduced the naturalistic Atari-style fMRI dataset and find that theory-based RL representations (Tsividis et al., 2021) in prefrontal cortex, showing that EMPA outperforms model-free baselines such as DDQN (Van Hasselt et al., 2016) in predicting frontal activity during gameplay. These studies demonstrate that features from RL agents map onto neural activity associated with decision-making and goal-directed control. However, because RL agents are optimized primarily for reward-based action learning, their representations tend to encode action-relevant features and may miss the richer perceptual, semantic, and predictive structure humans use to represent objects, dynamics, and future outcomes during gameplay.

World models, structured internal representations of the environment that support prediction and planning, were originally proposed in cognitive science (Ritchie, 1944; Johnson-Laird, 1983) and later formalized in deep-RL (Ha & Schmidhuber, 2018). Recent work has begun probing whether modern foundation models, particularly VLMs and LAMs, acquire analogous internal world models when interacting with games and GUIs (Waytowich et al., 2024; Wang et al., 2025a; Xie et al., 2026), but these evaluations focus on behavioral benchmarks (task success, generalization) rather than neural alignment. Our work bridges these two lines: we systematically evaluate the brain alignment of modern VLMs and LAMs on the same Atari-style fMRI dataset previously used to study theory-based RL (Tomov et al., 2023), comparing them directly against EMPA and DDQN baselines and probing how reasoning- and action-focused prompts shape brain-relevant representations across the cortical hierarchy.

B fMRI Preprocessing

Preprocessing pipeline. We preprocess the publicly available raw BIDS-formatted fMRI data using fMRIPrep v25.2.0 (Esteban et al., 2019; 2020). This performs the same core preprocessing stages as the original SPM12-based analysis of Tomov et al. (2023) such as, head-motion correction, coregistration, and normalization to MNI space, using a standardized, validated pipeline. We subsequently applied spatial smoothing with an 8-mm full-width-at-half-maximum (FWHM) Gaussian kernel, and included the estimated motion parameters as nuisance regressors in the encoding analysis.

Nuisance regression. Before surface projection, each MNI-space functional run is denoised using `nilearn.image.clean_img` with 14 nuisance regressors: the six rigid-body motion parameters, their six temporal derivatives, and mean cerebrospinal-fluid and white-matter signals obtained from fMRIPrep. Confound regression is performed jointly with linear detrending and high-pass filtering at (1/128) Hz ($(\mathrm{TR} = 2)$ s). This procedure controls for residual motion-related and physiological variance beyond the initial head-motion correction. The same denoised time series is used for all VLM, LAM, EMPA, and DDQN encoding models.

Surface projection and parcellation. The denoised MNI-space functional data are projected to the `fsaverage6` cortical surface using `nilearn.surface.vol_to_surf` (contributors, 2026). The pial and white-matter surfaces define depth-based sampling across the cortical ribbon, yielding 40,962 vertices per hemisphere. We remove the first and last five TRs from each run, leaving 273 TRs per run, and apply an `fsaverage6`-compatible Glasser parcellation for the ROI analyses.

Additional confound controls. We address game identity and temporal train-test leakage using complementary cross-validation scheme. Leave-one-run-out cross-validation evaluation holds out complete scanner runs and eliminates within-run adjacency between training and test TRs. Input windows are constructed separately within each run and never cross run boundaries.

We do not include separate button-press or task-specific visual-event regressors identical to those used by Tomov et al. (2023); therefore, we cannot completely exclude contributions from motor execution or task-specific visual events. As a complementary control, we evaluate low-level color-statistics and object/spatial scene-description prompts, which yield lower alignment than task-framed prompts containing goal and state-conditioned action information. However, these prompt controls do not replace explicit button-press regression. We therefore interpret frontal-motor findings as representational alignment and not as direct evidence of motor planning or execution.

C Detailed sub-ROIs of frontal-parietal, visual, and motor regions

The data covers eleven brain regions of interest (ROIs) in the human brain with the following subdivisions: (i) angular gyrus (AG): PFm, PGs, PGi, TPOJ2, TPOJ3, (ii) inferior frontal gyrus triangularis (IFGtriang): 44, 45, IFJa, IFSp, (iii) inferior frontal gyrus opercularis (IFGoperc): 6r, 6v, FOP1, FOP4, (iv) middle frontal gyrus (MFG): 55b, (v) inferior occipital gyrus (IOG): V8, VVC, VMV3, VMV2, VMV1, (vi) middle occipital gyrus (MOG): LO1, LO2, LO3, V3CD, MT, MST, FST, V4t, (vii) superior occipital gyrus (SOG): V7, V3a, V3b, V3cd, (viii) calcarine fissure (CAL): V1, V2, (ix) lingual gyrus (LING): VMV1, VMV2, VMV3, (x) cuneus (CUN): V3, POS1, V6, and supplementary motor area (SMA): SFL, 6ma, 6mp, SCEF (Baker et al., 2018; Milton et al., 2021; Desai et al., 2023). Fig. 6 shows flattened cortical surfaces for frontal-parietal, visual- and motor-selective regions displayed on the 'fsaverage' surface, used as the mask for all participants.

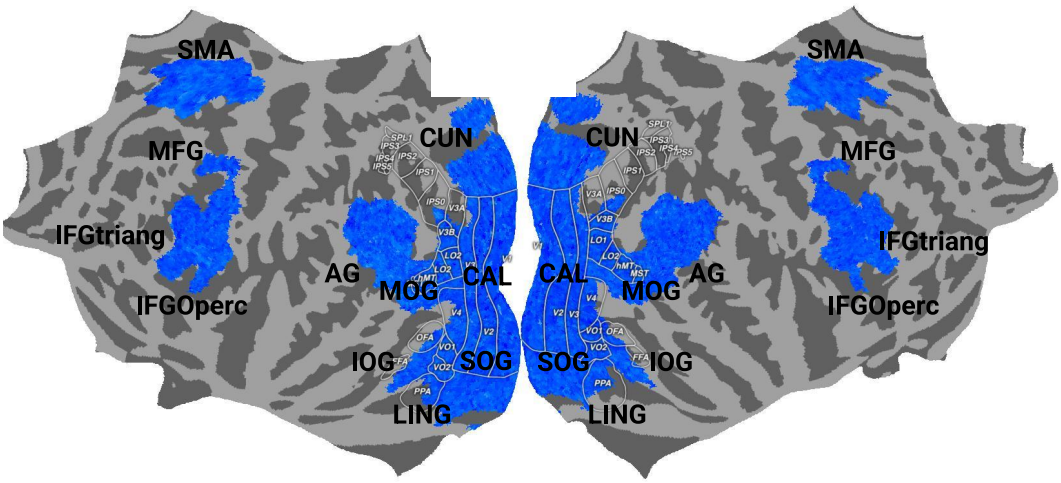


Figure 6: Flattened cortical surfaces for language-, visual- and motor-selective regions displayed on the 'fsaverage' surface, used as the mask for all participants.

D Baseline Models Features

We derived TR-aligned regressors from two computational models of Atari gameplay. For EMPA (Tomov et al., 2023), we extracted eight per-frame regressors. Four are continuous: *surprise* (the agent's prediction error on the observed next state), *spriteKL* (KL divergence between the agent's prior and updated beliefs about sprite dynamics, indexing belief revision over object behavior), *R_GG* (expected reward under the current goal-graph theory), and *num_effects* (count of distinct interaction effects observed at the current timestep). Four are binary indicators of theory revision: *replan_flag* (the agent re-planned its policy), *theory_change_flag* (the agent's theory of game rules changed), *newEffects_flag* (a previously-unseen interaction effect was observed), and *interaction_change_flag* (beliefs about an interaction between two object types changed). Concatenating these episodes within each level in instance order and averaging frame values within each TR over the level's expected start/end window, then trimming 5 TRs from each run edge to match fMRI preprocessing, yielding a (T, 8) feature matrix per subject. For DDQN (25M-step agent) (Van Hasselt et al., 2016), we loaded dqn_regressors_25M.bson and extracted per-episode action sequences

as (keycode, frame_idx, timestamp) tuples, mapping pygame keycodes to action labels and computing normalized action distributions per run for behavioral comparison against human and EMPA policies.

E Prompt Templates for Brain Encoding Models

Action Prompt

Action Plan:
"You are controlling the avatar in this game.
Look at the current game state.
What is your immediate next action and why?
Consider threats, goals, and optimal path."

Reasoning Prompt

Reasoning:
"Analyze this game step by step:
What objects do you see?
What are the spatial relationships?
What are the possible threats or goals?
What action would you recommend and why?"

Action Prompt (full)

Motor commitment + justification:
"You are controlling the avatar in this game.
Choose the movement key to press right now
and explain why it beats the other options open to you.
Answer briefly."

Action Prompt (bare)

Motor commitment, no justification:
"You are controlling the avatar in this game.
Name the single movement key you would press right now.
Give the key and nothing else at all.
Answer briefly."

Goal Prompt

Hazards/objectives, no motor commitment:
"You are looking at a frame from a game.
Identify which visible things are hazards that threaten the avatar
and which are objectives it must reach.
Answer briefly."

SCENE PROMPT (BASELINE)

Object semantics + spatial relations, no goals:
"You are looking at a frame from a game.
Describe the objects that are visible in it
and where each one sits relative to the others around it.
Answer briefly."

LOW-LEVEL PROMPT

Image-referring task, no game semantics or goals:
"You are looking at a still image.
Report which colours cover the largest areas of the picture
and roughly what fraction of it each one covers.
Answer briefly."

SCRAMBLED PROMPT (CONTROL – FAILED)

Text present, syntax destroyed, task lexicon retained:
"and this are press controlling to key options game the beats movement
why now the you in open right the explain You avatar it Choose to other.
Answer briefly."

SCRAMBLED PROMPT (NEUTRAL CONTROL)

Text present, syntax destroyed, task-irrelevant lexicon:
"a volumes room quiet library the and where bound shelves a cloth
spring windows has narrow fountain old since reading faded face The
hold courtyard in stood stone.
Answer briefly."

QWEN3.5 PROMPT

Reasoning prompt (with thinking enabled):
"Analyze this game frame step by step:
What objects/entities do you see?
What are the spatial relationships?
What are the possible threats or goals?
What action would you recommend and why?
IMPORTANT: Be brief. Keep your private reasoning under 800 tokens, then end the reasoning block and give a one-line final answer."

Prompt-dependent changes in model representations. We next examined whether the prompt conditions altered the model representations independently of their relationship with brain activity. For each condition, we constructed a representation matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ from the same $n = 126$ matched game levels. We decomposed its total energy into a condition-mean component, $n\|\boldsymbol{\mu}\|_2^2 / \|\mathbf{X}\|_F^2$, and a stimulus-varying component, $\|\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^\top\|_F^2 / \|\mathbf{X}\|_F^2$. We then used linear centered kernel alignment (CKA) to compare representational geometry across prompt conditions. Uncentered CKA retains the condition means, whereas centered CKA compares only the stimulus-varying components.

For the VLM, the condition mean accounted for 48.2%–97.3% of the total representational energy, depending on the prompt. Consequently, uncentered cross-condition CKA was uniformly high

Table 1: Prompt battery used to isolate the effects of text presence, visual and goal information, and motor commitment.

Condition	Axis	Isolates
`action`	Primary prompt	Original action prompt
`reasoning`	Primary prompt	Original reasoning prompt
`scrambled-neutral`	Text presence	Text present, syntax destroyed, and task-irrelevant lexicon
`scrambled`	Text presence	Text present, syntax destroyed, and task lexicon retained
`lowlevel`	Visual vs. goal	Colour statistics only, without game semantics
`scene`	Visual vs. goal	Object semantics and spatial relationships, without goals
`goal`	Action vs. reasoning	Hazards and objectives, without motor commitment
`action-bare`	Action vs. reasoning	Motor commitment only: name the movement key without justification
`action-full`	Action vs. reasoning	Motor commitment with justification relative to alternative actions

(0.857–0.995). After removing each condition mean, however, cross-condition CKA decreased to 0.028–0.173. A similar pattern was observed for the LAM, for which centered cross-condition CKA ranged from 0.050 to 0.251. Scene-based prompts retained the largest stimulus-varying component in the VLM (51.8%), whereas action-oriented representations were generally more mean-dominated; this was most pronounced for the LAM action-bare condition, for which the condition mean accounted for 98.4% of the total energy. These results show descriptively that prompts affect not only global representational offsets but also the geometry of stimulus-dependent variation. This analysis characterizes the model representations themselves and does not establish their correspondence with neural activity.

Table 2: **Energy decomposition of VLM representations across prompt conditions.** Values indicate the percentage of total embedding energy attributable to the condition-specific mean and to the remaining stimulus-varying component across 126 matched game levels.

Condition	Condition mean (%)	Stimulus-varying (%)
Action	97.3	2.7
Reasoning	72.2	27.8
Action-bare	89.4	10.6
Action-full	60.2	39.8
Goal	89.5	10.5
Scene	48.2	51.8
Low-level	60.9	39.1
Scrambled	60.8	39.2
Scrambled-neutral	79.1	20.9

Table 3: **Uncentered linear CKA between VLM representations.** Condition means are retained in this analysis. Uncentered CKA is uniformly high across conditions, consistent with the strongly mean-dominated representations reported in Table 2. Abbreviations: Act., Action; Reas., Reasoning; A-bare, Action-bare; A-full, Action-full; Low, Low-level; Scr., Scrambled; and Scr.-N, Scrambled-neutral.

	Act.	Reas.	A-bare	A-full	Goal	Scene	Low	Scr.	Scr.-N
Action	1.000	0.924	0.995	0.934	0.988	0.965	0.966	0.963	0.973
Reasoning	0.924	1.000	0.925	0.857	0.913	0.899	0.895	0.891	0.909
Action-bare	0.995	0.925	1.000	0.931	0.984	0.963	0.964	0.960	0.971
Action-full	0.934	0.857	0.931	1.000	0.921	0.908	0.910	0.913	0.913
Goal	0.988	0.913	0.984	0.921	1.000	0.953	0.955	0.951	0.961
Scene	0.965	0.899	0.963	0.908	0.953	1.000	0.936	0.936	0.943
Low-level	0.966	0.895	0.964	0.910	0.955	0.936	1.000	0.939	0.943
Scrambled	0.963	0.891	0.960	0.913	0.951	0.936	0.939	1.000	0.938
Scrambled-neutral	0.973	0.909	0.971	0.913	0.961	0.943	0.943	0.938	1.000

F Hyperparameter Details

We use the HuggingFace `transformers` interface to load the pretrained VLM and LAM models and their corresponding processors, and use 40-core GPU with 128GB-RAM with half precision (FP16) as compute.

Table 4: **Centered linear CKA between the stimulus-varying components of VLM representations.** Each condition-specific mean was removed before computing CKA. The low cross-condition similarities indicate that different prompts yield distinct stimulus-varying representational geometries. Abbreviations follow Table 3.

	Act.	Reas.	A-bare	A-full	Goal	Scene	Low	Scr.	Scr.-N
Action	1.000	0.053	0.071	0.058	0.087	0.108	0.077	0.076	0.042
Reasoning	0.053	1.000	0.046	0.029	0.041	0.099	0.061	0.047	0.028
Action-bare	0.071	0.046	1.000	0.045	0.028	0.074	0.051	0.054	0.036
Action-full	0.058	0.029	0.045	1.000	0.045	0.124	0.074	0.143	0.066
Goal	0.087	0.041	0.028	0.045	1.000	0.141	0.095	0.073	0.034
Scene	0.108	0.099	0.074	0.124	0.141	1.000	0.170	0.173	0.085
Low-level	0.077	0.061	0.051	0.074	0.095	0.170	1.000	0.129	0.036
Scrambled	0.076	0.047	0.054	0.143	0.073	0.173	0.129	1.000	0.070
Scrambled-neutral	0.042	0.028	0.036	0.066	0.034	0.085	0.036	0.070	1.000

We used bootstrap ridge-regression with the following parameters: MSE loss function, and L2-decay (λ) varied from 10^1 to 10^3; best λ was chosen by tuning on validation data that comprised a randomly chosen 10% subset from train set used only for hyper-parameter tuning.

G Details of explained variance partitioning

Using a variance partitioning approach (de Heer et al., 2017; LeBel et al., 2021), we test whether action- and reasoning-based prompts carry overlapping or distinct brain-relevant information, and whether LAMs, which are fine-tuned on action-specific environments reshapes this decomposition. In particular, we decomposed each model's joint R^2 into three components: variance shared between action and reasoning prompts, variance uniquely explained by action prompt, and variance uniquely explained by reasoning prompt (Vaidya et al., 2022; Deniz et al., 2019). For each subject, we fit three voxel-wise encoding models: one using the concatenated action and reasoning representations as joint predictors (yielding R^2_{joint}), one using only action representations (R^2_{action}), and one using only reasoning representations ($R^2_{reasoning}$). We then computed each component's contribution following standard variance-partitioning formulations: $R^2_{joint} - R^2_{reasoning}$ for unique action, $R^2_{joint} - R^2_{action}$ for unique reasoning, and $R^2_{action} + R^2_{reasoning} - R^2_{joint}$ for shared variance. We performed this analysis separately for two VLMs (Qwen2.5-VL, and InternVL3-8B) and LAMs (UI-TARS-7B-DPO and OS-Atlas-Pro-7B) on all 32 subjects, and evaluated each component at both whole-brain and ROI levels.

H Statistical significance

To determine if normalized predictivity scores are significantly higher than chance, we use block permutation tests. We employ the standard implementation of a block permutation test for fMRI data, which is to split the fMRI data into blocks of 10 contiguous TRs and permute the order of these blocks, while maintaining the original order of the TRs within each block. By permuting predictions 5000 times, we create an empirical distribution for chance performance, from which we estimate the p-value of the actual performance. To estimate the statistical significance of performance differences, such as between the model's predictions and chance, we utilized the Wilcoxon signed-rank test, applying it to the mean correlations for the participants. In all cases, we denote significant differences with an asterisk *, indicating cases where p≤ 0.05.

I Detailed Feature Extraction

Let $\mathbf{x}_{t,j}$ denote the j-th RGB game frame sampled for fMRI TR t (converted to a PIL image), and let p denote the text prompt (e.g., action or reasoning; empty in the *no-prompt* condition). Rather than conditioning on a single frame, we give the model a short multi-frame window for each TR. The game runs at $\sim$20 FPS, so each 2-s TR spans $\sim$40 frames; we sample $k = 4$ frames evenly $\sim$10 frames ($\approx$0.5 s) apart, ending at the last frame of the TR:

$$\mathbf{X}_t = \big(\mathbf{x}_{t,1}, \mathbf{x}_{t,2}, \mathbf{x}_{t,3}, \mathbf{x}_{t,4}\big). \tag{1}$$

Frames are taken from each participant's own gameplay, and windows are constructed within each run, never crossing run boundaries. We build a chat-style input in which the k frames are passed as

an ordered sequence of image items followed by the prompt:

$$\text{input}_t = \big[\texttt{<image>}{=}\mathbf{x}_{t,1};\ \ldots;\ \texttt{<image>}{=}\mathbf{x}_{t,4};\ \texttt{<text>}{=}p\big], \tag{2}$$

which is serialized with each model's chat template (with the generation marker added) and tokenized by the model-specific processor. This uses each model's native multi-image input and gives the model access to short-term motion cues rather than a single static observation.

Generation and hidden states. We run `model.generate()` with greedy decoding, and hidden-state outputs enabled. Let L denote the number of transformer blocks and d the hidden dimension. At each decoding step, the model returns hidden states for all $L + 1$ layers (embedding layer and L transformer blocks). We take the hidden states at the final decoding step, i.e., the representation from which the last output token is predicted, excluding the end-of-sequence token. Because generation length varies across TRs, this readout is taken at a variable position. For layer $\ell \in \{0, \ldots, L\}$, let $\mathbf{e}_t^{(\ell)} \in \mathbb{R}^d$ denote this final-step hidden state. Because of causal self-attention over the full context, $\mathbf{e}_t^{(\ell)}$ aggregates information from all k frames, the prompt, and the generated response. We concatenate the per-layer vectors into

$$\mathbf{E}_t = \big[\mathbf{e}_t^{(0)};\ \mathbf{e}_t^{(1)};\ \ldots;\ \mathbf{e}_t^{(L)}\big] \in \mathbb{R}^{(L+1)\times d}. \tag{3}$$

Run-level features. For a run with T TRs, we compute one feature tensor per TR and stack them into

$$\mathbf{E} = \{\mathbf{E}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times(L+1)\times d}. \tag{4}$$

These tensors serve as predictors for the voxel-wise encoding models, enabling layer-wise analyses and placing VLM and LAM features on the same TR-level footing as the EMPA regressors.

J Behavioral Analysis

Model–human action agreement. Gameplay logs are resampled at 2 Hz (every tenth frame of the ≈20 Hz logs), so agreement is evaluated at a finer temporal resolution than the TR-level encoding features. At each sampled frame, we give the model the action prompt (App. E), which asks for its immediate next action, together with a four-frame window constructed as in the encoding analysis but ending at that frame. We extract one action from the generated text with a parser fixed before inspecting any model output: a closed alias table maps 30 surface forms onto the six canonical actions (`LEFT`, `RIGHT`, `UP`, `DOWN`, `SPACEBAR`, `NOOP`), including semantic variants ("fire"/"shoot"→`SPACEBAR`; "do nothing"/"wait"→`NOOP`). A bare action word in first position takes precedence; otherwise, the earliest whole-word alias in the response is used. Responses containing no alias are treated as parse failures and excluded from the agreement computation rather than scored as disagreement; parse-failure rates (0–2.3%) are reported in Table 5.

The participant's action at a sampled frame is the key held at that timestamp (each key-down remains active until the next, and the final one for 1 s); if none is held, we use the nearest key-down within 0.5 s, and otherwise drop the frame. Agreement is the fraction of matched frames with identical actions, computed per participant and game, and compared with two baselines: uniform chance over the actions the participant used ($1/k$, $k = 4$–5) and the majority-class rate (always predicting the participant's most frequent action). Because action distributions are skewed toward no-ops and repeated moves, we report Cohen's κ, macro-F_1, and balanced accuracy as primary statistics (Table 5); raw accuracy against uniform chance overstates agreement and is included only for comparability.

Table 5: Model–human action agreement across games (mean ± SD across 32 participants). Chance denotes the uniform-chance level for the available actions in each game.

Game	Chance	VLM agreement	LAM agreement
bait	25.0%	$24.5 \pm 9.1\%$	$27.2 \pm 8.3\%$
chase	25.0%	$21.9 \pm 14.3\%$	$30.5 \pm 13.9\%$
helper	25.0%	$28.1 \pm 12.4\%$	$31.3 \pm 13.0\%$
lemmings	20.0%	$29.7 \pm 6.3\%$	$26.9 \pm 5.7\%$
plaqueAttack	20.0%	$20.4 \pm 1.0\%$	$17.7 \pm 3.1\%$
zelda	20.0%	$35.2 \pm 12.3\%$	$36.1 \pm 13.5\%$

Model–human action agreement. To contextualize the representational findings with participants' behavior, we compare the actions recommended by the VLM and LAM with the actions performed by each participant. As shown in Table 5, agreement is modest and varies across games. Both model families are numerically above chance for *helper*, *lemmings*, and *zelda*, with the highest agreement observed for *zelda*. However, several model–game combinations remain near or below chance, including the VLM on *bait* and *chase* and the LAM on *plaqueAttack*. Overall, neither model family shows a consistent advantage across games. Moreover, because the models process observations offline without executing actions or receiving subsequent states and rewards, closed-loop measures such as reward, win rate, and task success are not defined in the present setting. These results therefore provide behavioral context for the neural alignment findings but do not establish that either model reproduces the participants' action policy.

Action decoding. We test whether participants' actions can be decoded from model representations during gameplay. For each TR, the target is the participant's action in that TR, defined as the most frequent key within the TR (a TR-level analogue of the frame-level definition used for agreement). The predictors are the same TR-level features used in the encoding analysis: the final-step hidden states of the action-prompt generation at the layer selected for encoding (VLM L18, LAM L21); for EMPA, we use its eight TR-level regressors. We train an L2-regularized logistic regression, with regularization strength selected by inner cross-validation, per participant with leave-one-run-out cross-validation, so that no test TR shares a run with the training data. Because the action distribution is imbalanced, we report balanced accuracy as the primary metric, with chance defined per participant as $1/k$ for the k actions present in that participant's data (mean chance = 0.167) (Table 6).

Decoding participants' actions from model representations. We next test whether the model representations contain information predictive of participants' actions. Because the model representations are constant within each game level, the prediction target is each participant's dominant action for that level; this is therefore a level-wise rather than TR-wise decoding analysis. Using the layers selected in the brain-encoding analysis, both model families exceed the six-action chance floor of 0.167 under leave-one-run-out cross-validation (Table 6). Balanced accuracy reaches 0.342 ± 0.083 for the VLM and 0.362 ± 0.088 for the LAM, compared with 0.289 ± 0.073 for EMPA. However, raw accuracy is only slightly higher than the majority-action baseline for the VLM (0.503 vs. 0.475) and LAM (0.512 vs. 0.475). Moreover, the LAM exceeds the VLM by only 0.020 in balanced accuracy, indicating a small numerical difference rather than a reliable action-tuning advantage. Thus, the model representations contain modest level-wise information about participants' actions, but these results do not establish strong behavioral prediction or human-like policy use.

Table 6: Decoding participants' dominant level-wise actions from model features using leave-one-run-out cross-validation. Values are mean ± SD across 32 participants. Balanced accuracy is compared with the six-action chance floor of 0.167.

Model	Layer	Balanced acc.	Δ chance	Raw acc.	Majority baseline
VLM	L18	0.342 ± 0.083	$+0.175$	0.503	0.475
LAM	L21	0.362 ± 0.088	$+0.195$	0.512	0.475
EMPA	N/A	0.289 ± 0.073	$+0.122$	0.280	0.460

This action decoding analysis provide exploratory evidence that representations aligned with brain activity also carry TR-level information about participants' actions. We interpret them cautiously: they link brain-aligned features to behavior at the representational level but do not establish strong behavioral prediction, human-like policy use, or competent gameplay.

K GENERALIZATION ACROSS MODEL FAMILIES

To verify that the prompted-alignment effect is not specific to a single VLM or LAM, we replicate the whole-brain analysis on a second model from each family: InternVL3-8B (VLM) and OS-Atlas-Pro-7B (LAM). Fig. 7 shows that all four models achieve substantial brain alignment under both reasoning and action prompts, with all four significantly exceeding the RL baselines (EMPA, DDQN). Within the VLM family, both Qwen2.5-VL (reasoning $r = 0.050$, action $r = 0.044$) and InternVL3 (reasoning $r = 0.049$, action $r = 0.045$) show a small reasoning-leaning pattern. Within the LAM family, both UI-TARS (reasoning $r = 0.046$, action $r = 0.050$) and OS-Atlas (reasoning $r = 0.042$, action $r = 0.044$) show a small action-learning pattern. The within-model differences

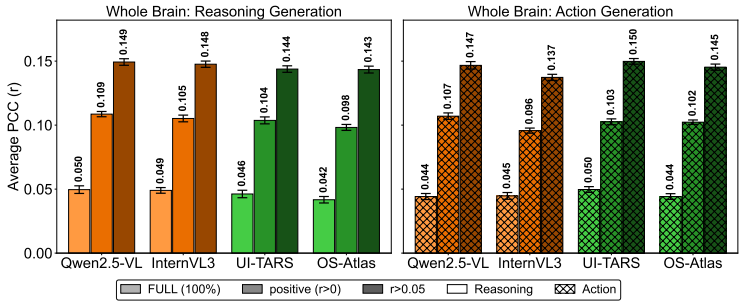


Figure 7: **Whole-brain alignment across two VLMs and two LAMs under reasoning and action prompts.** Average Pearson correlation across participants and voxels per model, under three voxel-selection criteria: all voxels (Full), positively predicted (r>0) and (r>0.05). Left: reasoning prompts. Right: action prompts (hatched). Orange: VLMs; Green: LAMs. Darker shades denote the primary model in each family (Qwen2.5-VL, UI-TARS); lighter shades denote the secondary model (InternVL3, OS-Atlas). Error bars denote mean ± SEM across subjects.

are small (≤ 0.006) and do not reach significance, so we treat this as a directional pattern rather than an established prompt preference; For the remainder of the paper, we focus our ROI-level (RQ2) and variance-partitioning (RQ3) analyses on Qwen2.5-VL and UI-TARS as representative members of each family; corresponding results for InternVL3 and OS-Atlas are reported in App. N.

L Prompt Controls

We next examine whether the observed prompt effects can be explained by text presence, visual description, goal information, or motor commitment alone (see Table 1). In the primary unthresholded analysis, as shown in Fig. 8, (left), the original action and reasoning prompts produce consistently high brain alignment for both model families and exceed the EMPA baseline. The action-full condition also shows relatively high alignment (≈ 0.040 and 0.049), whereas action-bare is model-dependent, producing higher alignment for the LAM (≈ 0.055) than for the VLM (≈ 0.029). The goal condition yields intermediate and model-dependent performance (≈ 0.040 for the VLM, 0.029 for the LAM). In contrast, scrambled text, low-level colour descriptions, and scene descriptions generally produce weaker alignment than the original task-framed prompts ($r \approx 0.006$–0.029), although the magnitude varies between the VLM and LAM. The same broad pattern is visible in the $r > 0$ and $r > 0.05$ sensitivity analyses. These results indicate that text presence or visual description alone is insufficient to reproduce the consistently stronger alignment elicited by task-instruction prompts. However, because the control effects are model-dependent, we do not interpret them as demonstrating that each prompt isolates a single action- or reasoning-specific semantic dimension.

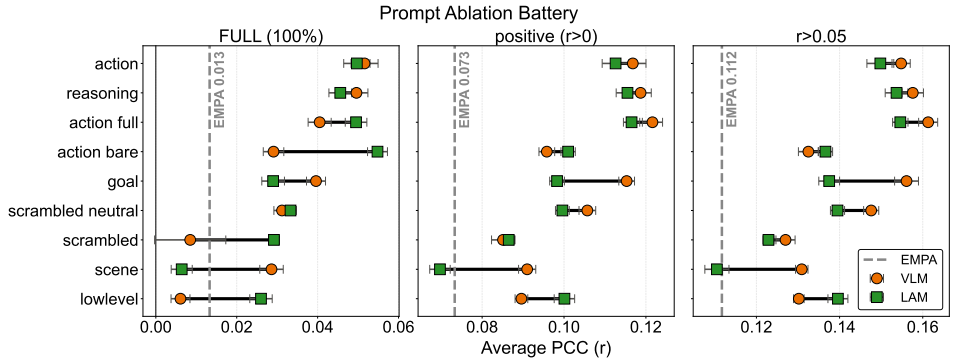


Figure 8: **Prompt-control battery.** Whole-brain alignment for the VLM (Orange circles, Qwen2.5-VL) and LAM (Green squares, UI-TARS) under the two primary prompts and the seven control prompts of Table 1, shown for all cortical vertices (FULL, primary metric), positively predicted vertices ($r > 0$), and vertices with $r \geq 0.05$. Dashed vertical line: EMPA under the same criterion. Error bars: mean ± SEM across $N = 32$ subjects.

M VLM AND LAM: PROMPT-DRIVEN ALIGNMENT GAINS ACROSS ROIS

As shown in Fig. 9, both VLM and LAM representations outperform the classical baselines in whole-brain voxel-wise encoding without any prompt instruction (shown for voxels with $r \geq 0.05$). Specifically, VLM (*no-prompt*) achieves r=0.131 and LAM (*no-prompt*) reaches r=0.127, compared to EMPA (r=0.112) and DDQN (r=0.094). The advantage is consistent across all three feature dimensionalities tested. Crucially, even when VLM and LAM features are reduced to 8 dimensions to match EMPA's feature count, they continue to significantly show higher brain alignment than both RL baselines (VLM r=0.122, LAM r=0.119), ruling out feature count as the source of the gain. Performance saturates between 64-dim and 1024-dim, suggesting that the gain over RL baselines reflects what the features encode rather than feature dimensionality alone. Together, these results indicate that contemporary multimodal architectures learn richer latent representations of dynamic game environments than traditional task-optimized RL systems.

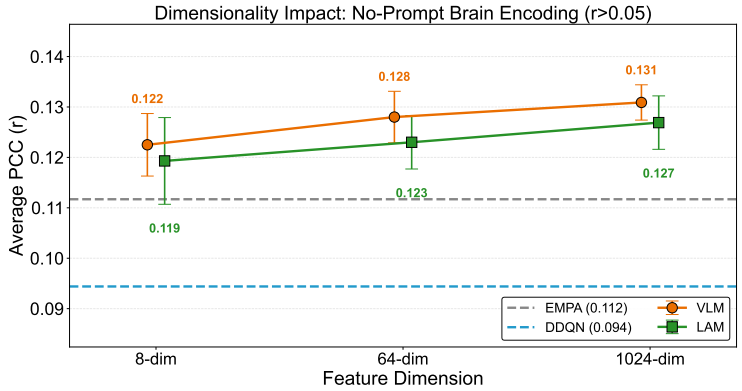


Figure 9: **Whole-brain voxel-wise encoding performance.** Effect of feature dimensionality under the *no-prompt* condition, shown for voxels with $r \geq 0.05$:: VLM (orange) and LAM (green) features evaluated at three dimensionalities (8, 64, 1024), compared against the RL baselines EMPA (gray dashed, r=0.112) and DDQN (light blue dashed, r=0.094). * indicates significant pairwise gain from 8 to 64 dim (paired t-test, $p < 0.05$); the 64→1024 transition is non-significant, indicating saturation. At every dimensionality tested, both VLM and LAM significantly exceed both RL baselines. Error bars: mean ± SEM across subjects.

We repeat the ROI analysis on LAMs under Action, Reasoning, and *no-prompt* settings (Fig. 14). As with VLMs, prompting improves alignment in every ROI ($q < 0.05$, FDR-corrected). Gains are largest in SMA (reasoning/action $r = 0.075/0.078$ vs. *no-prompt* $r = 0.020$; $\Delta r = +0.057$, $q < 0.001$) and MFG ($0.060/0.071$ vs. 0.024; $\Delta r = +0.042$, $q < 0.001$), followed by IOG and IFGtriang (both $\Delta r = +0.030$), SOG and AG (both $+0.023$), LING ($+0.021$), and MOG ($+0.019$), and smallest in IFGoperc ($+0.016$), CUN ($+0.010$), and CAL ($+0.009$). Within the LAM, action-prompt alignment is numerically equal to or higher than reasoning-prompt alignment in 10 of 11 ROIs, with IFGtriang the only exception (0.047 vs. 0.046), whereas the VLM shows a mixed pattern (Fig. 3). These within-model action–reasoning differences are small and not individually significant; we examine them with variance partitioning in the main text (Sec. 5, RQ3).The same pattern holds when restricting to vertices with $r > 0$ or $r > 0.05$ (Fig. 14, middle and bottom panels), where all 11 ROIs again show significant prompt gains.

Table 7: **Peak model layers across cortical ROI groups for the reasoning condition.** Each entry reports the peak layer and its group-mean held-out Pearson correlation in parentheses, using all vertices and $N = 32$ participants. Ties indicate equal values at three-decimal precision. Complete layer profiles are provided in Appendix Fig. 11 and 16.

Model	Early visual	Higher visual	Motor/planning	Prefrontal	Parietal association
VLM	L4 (0.028)	L0/L4 (0.016)	L18 (0.067)	L18 (0.061)	L28 (0.025)
LAM	L7 (0.023)	L7 (0.020)	L7/L18 (0.052)	L21 (0.036)	L18 (0.015)

Layer-wise regional preferences. We next examine which model layers provide the highest unthresholded encoding performance within each cortical group (Table 7). For the VLM reasoning condition, the early- and higher-visual groups peak at L4 and L0/L4, respectively, whereas mo-

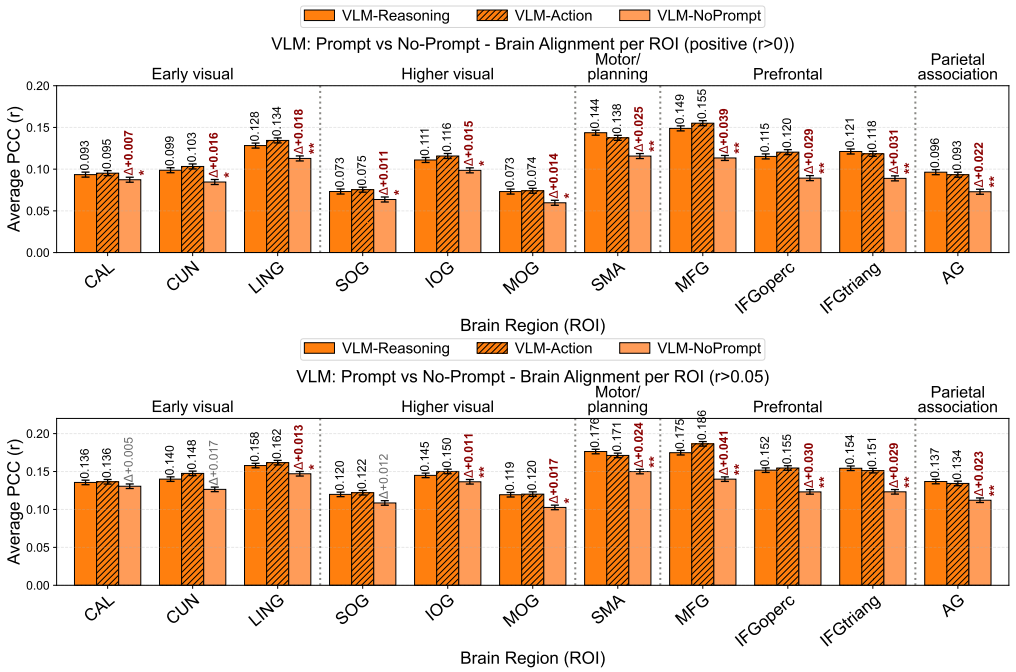


Figure 10: Brain alignment averaged across participants and voxels per ROI for prompted vs. no-prompt VLM representations. Average Pearson correlation between predicted and observed fMRI responses across 11 ROIs for VLM-Reasoning (solid), VLM-Action (hatched), and VLM-NoPrompt (light). Error bars denote mean ± SEM across subjects. Markers denote paired comparisons against no-prompt (R: Reasoning, A: Action; *p $<$ 0.05, **p $<$ 0.01, ***p $<$ 0.001). Both prompted conditions show higher alignment than no-prompt across all ROIs.

tor/planning, prefrontal, and parietal regions peak at L18, L18, and L28. For the LAM, the early- and higher-visual groups both peak at L7, while motor/planning peaks jointly at L7/L18, prefrontal at L21, and parietal cortex at L18. Thus, higher-order association regions generally attain their peak alignment at deeper layers than visual regions. Because these peaks are descriptive and were selected from the evaluated layers, we do not interpret them as evidence of a statistically established monotonic hierarchy. Complete layer profiles are provided in Appendix Figs. 11 (VLM) and 16 (LAM).

N Variance Partitioning Results

Supplementary Fig. 19 extends the variance partitioning from Fig. 4 to the remaining 9 ROIs. Three patterns hold consistently across regions.

The reasoning-redundancy asymmetry scales with the cortical hierarchy. The LAM's unique-reasoning variance is negative in all seven ROIs (-6% to -14% of joint explained variance), whereas the VLM's remains positive throughout (4%–15%). The asymmetry is largest in frontal-motor cortex (SMA: -14%; prefrontal: -10%) and MOG (-10%), intermediate in SOG (-9%), parietal (-8%), and IOG (-7%), and smallest in early visual cortex (-6%). The VLM–LAM difference in unique-reasoning variance is significant in MOG, IOG, and parietal cortex ($p < 0.05$) and marginal in the remaining ROIs ($p < 0.10$). The magnitude therefore varies across regions but does not follow a strictly monotonic cortical ordering.

LAM unique-action variance is consistently 2–$5\times$ larger than VLM's (22%–34% vs. 5%–13% across ROIs), following the same hierarchical ordering: SMA (34%) SMA (34% vs. 13%) and is lowest in early visual cortex (22% vs. 10%), so the largest action-unique and most negative reasoning-unique components both occur in SMA. The action-uniqueness ordering thus mirrors the reasoning-redundancy ordering, both peaking in motor-planning cortex.

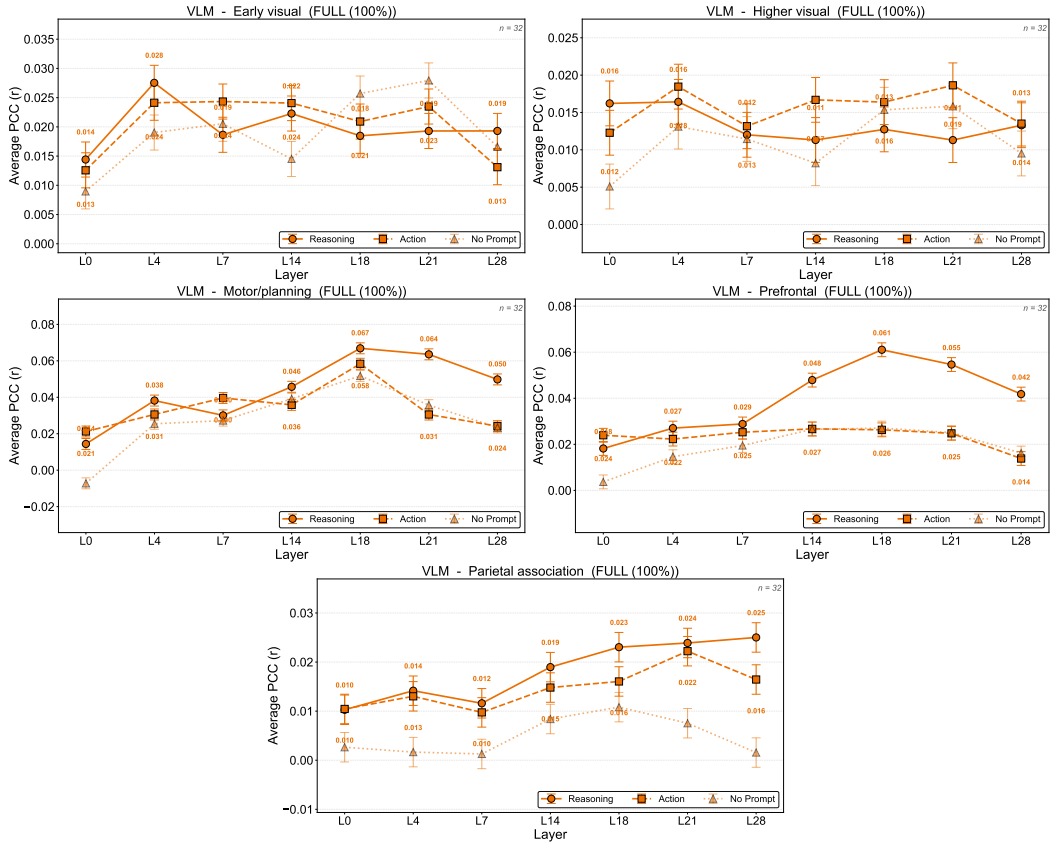


Figure 11: **Layer-wise brain alignment of VLM (Qwen2.5-VL) representations, all vertices (FULL).** Average Pearson correlation between predicted and observed fMRI responses at sampled layers (L0–L28) for reasoning (solid circles), action (dashed squares), and *no-prompt* (dotted triangles) representations, shown separately for five ROI groups: early visual (CAL, CUN, LING), higher visual (SOG, IOG, MOG), motor/planning (SMA), prefrontal (MFG, IFGoperc, IFGtriang), and parietal association (AG). Error bars: mean $\pm$ SEM across $N = 32$ participants. Visual groups peak at early layers, whereas motor/planning, prefrontal, and parietal groups peak at intermediate-to-deep layers (Table 7)

Shared variance dominates in both models (74%–85% of joint explained variance across all ROIs), indicating that most prompt-driven brain prediction reflects representations the two prompts engage in common. The VLM/LAM dissociation therefore does not lie in this shared bulk but in how each model allocates the residual prompt-unique variance: VLM splits it roughly evenly between action and reasoning, while LAM concentrates it on action, with reasoning becoming actively redundant in higher-order cortex.

N.1 Generalization across VLM–LAM pairs for variance partitioning: OS-Atlas-Pro vs InternVL3

To test whether the action-dominance vs prompt-symmetric dissociation observed within the Qwen-VL family (Section 5) generalizes to other architectures, we ran variance partitioning, comparing OS-Atlas-Pro-7B (Qwen2-VL-based LAM) and InternVL3-8B (Qwen2.5-VL-based VLM, cross-family). Because OS-Atlas and InternVL3 differ in architecture family, this comparison cannot isolate fine-tuning effects from architectural effects. We therefore treat it as supporting evidence of generalization across VLM–LAM pairs rather than a controlled within-family contrast. Figure 20 shows the explained variance across whole-brain, AG, and SMA. The qualitative dissociation is preserved across all three regions:

- **Whole-brain:** OS-Atlas shows strong action-dominance ($u_{\text{Action}} = 0.0102$, $u_{\text{Reasoning}} = 0.002$); InternVL3 is approximately prompt-symmetric ($u_{\text{Action}} = 0.004$, $u_{\text{Reasoning}} = 0.006$).

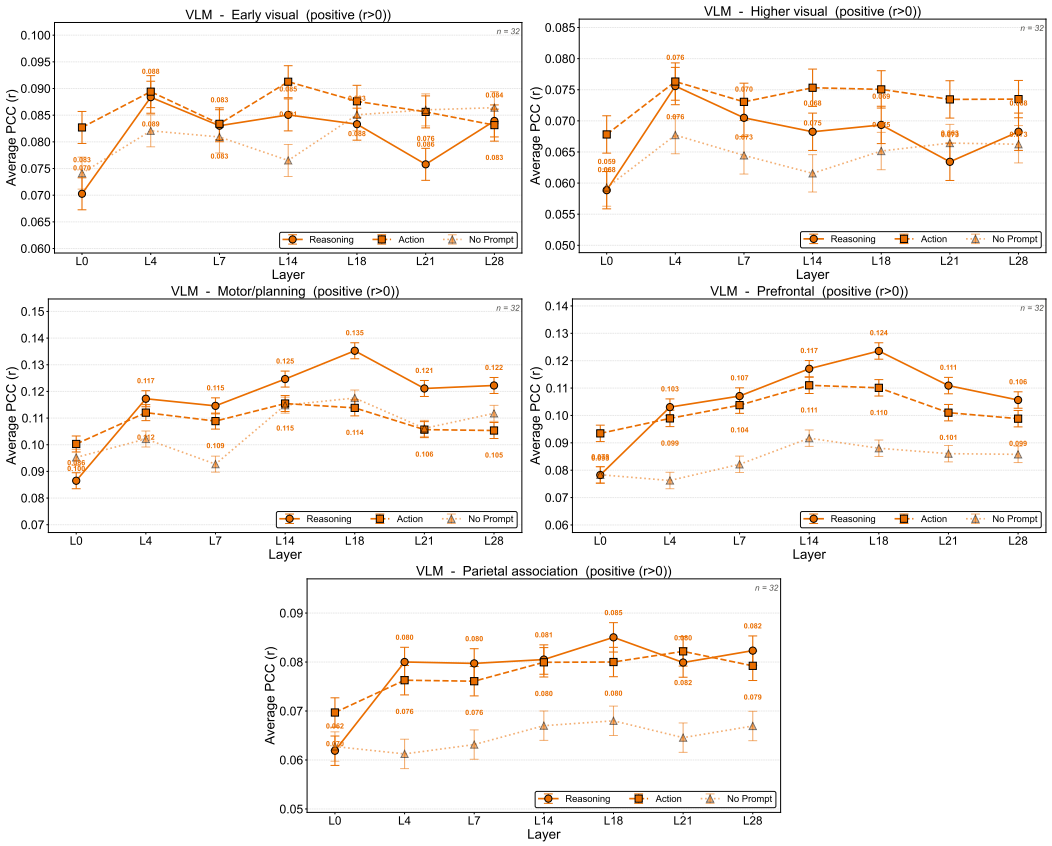


Figure 12: **Layer-wise brain alignment of VLM (Qwen2.5-VL) representations, positively predicted vertices** ($r > 0$)**.** Same as Fig. 11, restricted to vertices with $r > 0$. Error bars: mean $\pm$ SEM across $N = 32$ participants.

- **AG:** OS-Atlas action-asymmetric (0.009 vs. 0.003); InternVL3 Reasoning-leaning (0.002 vs. 0.007).
- **SMA:** OS-Atlas action-asymmetric (0.011 vs. 0.006); InternVL3 Reasoning-leaning (0.007 vs. 0.010).

The shared component is comparable between the two models in all three regions (0.031-0.037 at whole-brain, 0.018 to 0.021 at AG and SMA), indicating both models capture similar overlapping information. The dissociation manifests in the unique-variance components: action-tuning preferentially carves Action-relevant structure into the LAM representation, while the VLM's representation distributes variance more evenly between prompt conditions or shifts toward Reasoning in higher-order cortex. The cross-family preservation of the dissociation supports the interpretation that the within-family Qwen2-VL pattern reflects a general property of action-tuned vs general-purpose multimodal foundation models, not a quirk specific to the Qwen2-VL $\rightarrow$ UI-TARS fine-tuning step.

O [Additional analysis]: Alignment using generated reasoning traces of thinking-mode models.

Beyond prompt-conditioned representations from VLMs and LAMs (RQ1-RQ3), we conduct a preliminary study to examine whether the *generated* reasoning traces of thinking-mode models align with brain activity. Recent vision-language models such as Qwen3.5 (Qwen Team, 2026), positioned as "native multimodal agents," produce explicit reasoning sequences between `<think>` and `</think>` tags before generating a final answer. We exploit this thinking-mode capability to extract two distinct embeddings for each gameplay frame context: one from the reasoning span and one from the action output (see the Qwen3.5 prompt template in App. E.

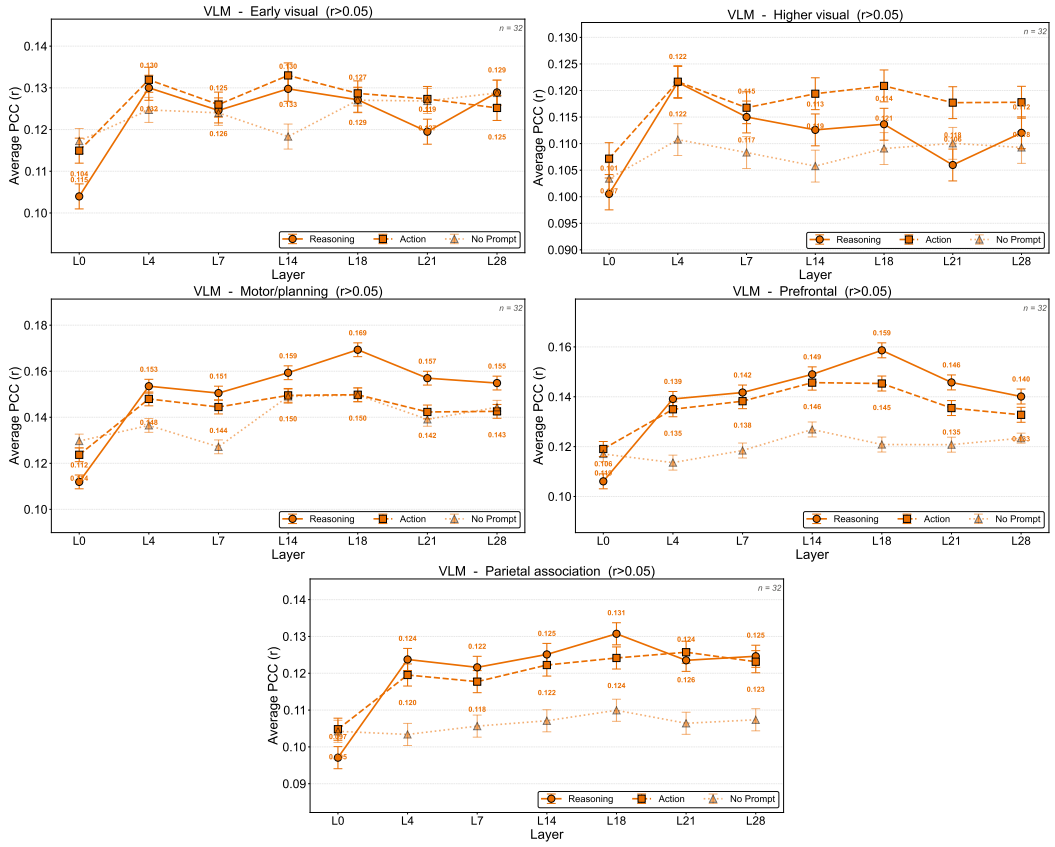


Figure 13: **Layer-wise brain alignment of VLM (Qwen2.5-VL) representations, vertices with $r > 0.05$.** Same as Fig. 11, restricted to vertices with $r > 0.05$. Error bars: mean $\pm$ SEM across $N = 32$ participants.

Specifically, we run generation with thinking enabled and extract two readouts from the resulting sequence: (i) the last-token hidden state at the end of the reasoning trace (before `</think>`), referred to as the *Reasoning last-token* readout; and (ii) the last-token hidden state at the end of the full generation (after `</think>`), referred to as the *Final-answer last-token readout*. To enable a consistent internal comparison, we apply the same generation-based feature extraction procedure used for Qwen2.5-VL and UI-TARS. Note that this comparison places Qwen3.5's generated reasoning content alongside VLM/LAM's prompt-conditioned encodings, related but distinct measurements.

Across subjects, we observe a substantial asymmetry in Qwen3.5. The final-answer readout (r=0.031 ± 0.005) shows higher brain alignment than the Reasoning last-token readout (r=0.012 ± 0.005), paired t-test $p = 0.032$, Wilcoxon $p = 0.043$. For reference, prompt-conditioned VLM and LAM achieve r$\approx$0.047 for both Action and Reasoning prompts, indicating that Qwen3.5's final answer output is comparably brain-aligned to prompt-conditioned multimodal models, while its reasoning trace is substantially weaker. To further examine whether this gap in Qwen3.5 reasoning alignment reflects the content of the reasoning trace or merely the readout position, we compare mean-pooled readouts across the same spans. The Qwen3.5 Reasoning mean-pooled readout improves to r=0.036 ± 0.010, reaching the level of the final-answer readout ($r = 0.031$), but remains substantially lower than the reasoning prompt-conditioned representations of VLMs and LAMs (r$\approx$0.047). The persistence of this gap under matched pooling suggests that the asymmetry is not solely a readout-position artifact. Why explicit reasoning traces show weaker brain alignment than the model's eventual action commitment - whether due to the content of the reasoning trace, residual positional effects, or trace-length variability - remains an open question. Together, these observations document a substantial reasoning-action asymmetry in thinking-mode brain alignment and motivate further investigation of how reasoning-trace structure interacts with brain alignment in such models.

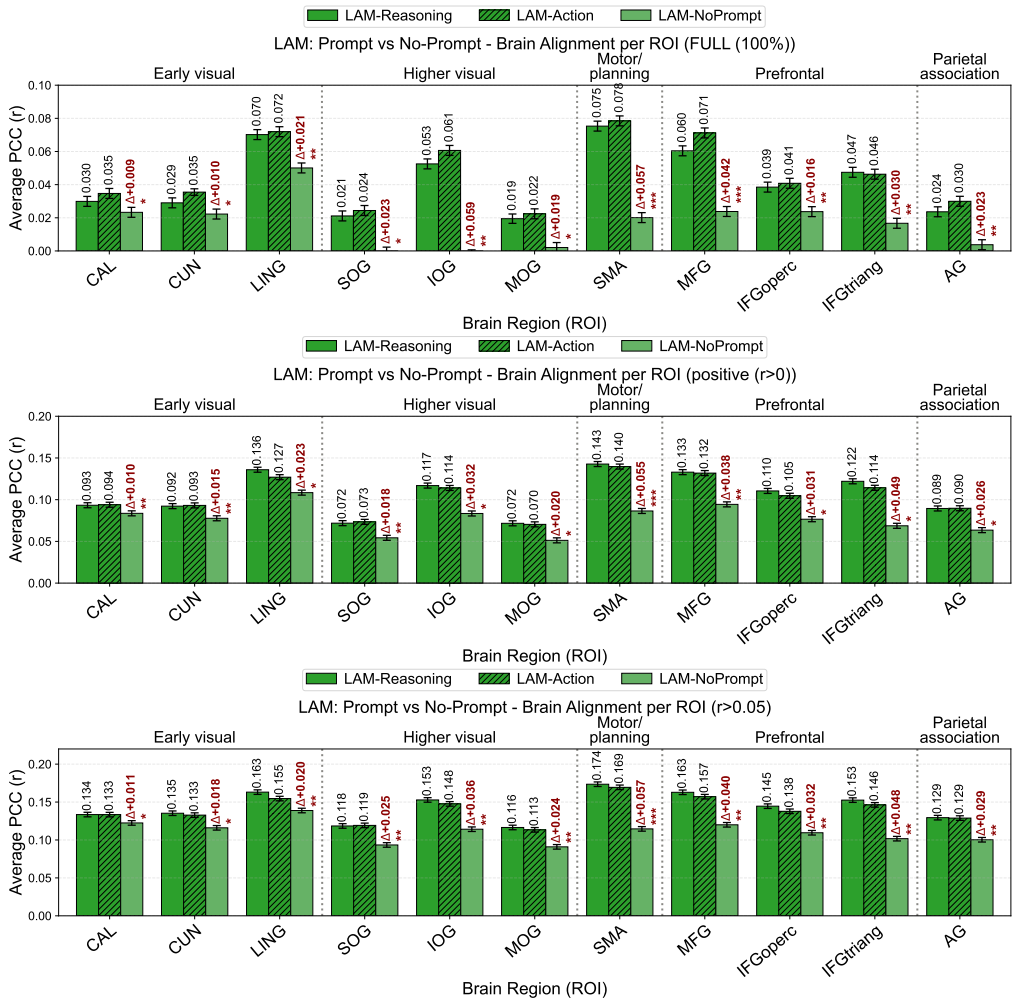


Figure 14: Brain alignment averaged across participants and voxels per ROI for prompted vs. no-prompt LAM representations. Average Pearson correlation between predicted and observed fMRI responses across 11 ROIs for LAM-Reasoning (solid), LAM-Action (hatched), and LAM-NoPrompt (light). Error bars denote mean ± SEM across subjects. Markers denote paired comparisons against no-prompt (R: Reasoning, A: Action; *p $<$ 0.05, **p $<$ 0.01, ***p $<$ 0.001). Both prompted conditions show higher alignment than no-prompt across all ROIs.

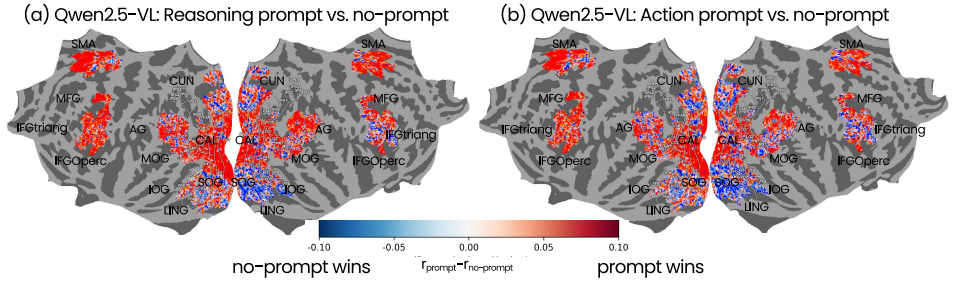


Figure 15: **Voxel-wise difference in brain alignment between prompted and *no-prompt* VLM representations.** Group-averaged across subjects cortical flatmap (both hemispheres, 'fsaverage') showing per-voxel differences in Pearson correlation, r_{prompted}-$r_{\text{no-prompt}}$, for Qwen2.5-VL under (a) reasoning prompts and (b) action prompts. Red voxels indicate higher alignment under the prompted condition; blue voxels indicate higher alignment under no-prompt.

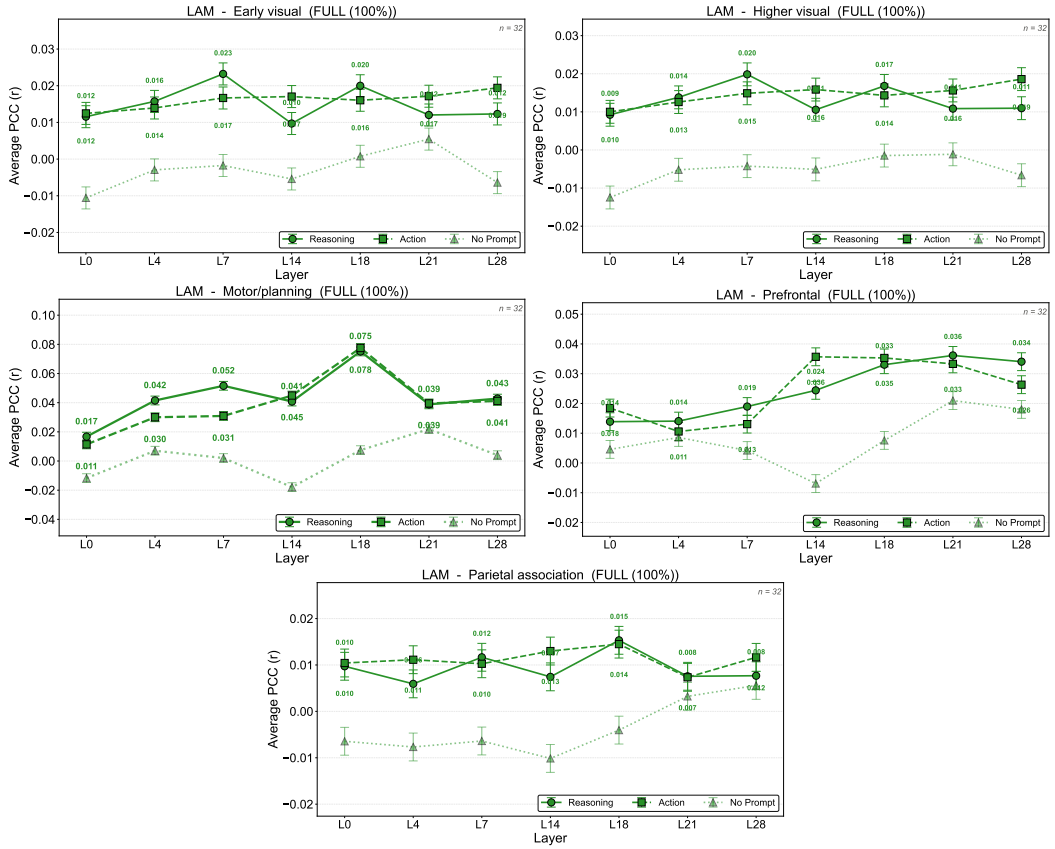


Figure 16: **Layer-wise brain alignment of LAM (UI-TARS) representations, all vertices (FULL).** Average Pearson correlation between predicted and observed fMRI responses at sampled layers (L0–L28) for reasoning (solid circles), action (dashed squares), and *no-prompt* (dotted triangles) representations, shown for the same five ROI groups as Fig. 14. Error bars: mean ± SEM across $N = 32$ participants. Prompted representations exceed *no-prompt* representations at all layers in every ROI group, and higher-order groups peak at intermediate-to-deep layers (Table 7).

P LIMITATIONS

Our study has several limitations that scope its claims and motivate future work. One possible limitation our study lies in matched-architecture VLM-vs-LAM comparison is restricted to the Qwen-VL family (Qwen2.5-VL; UI-TARS-7B-DPO and OS-Atlas-Pro-7B, both Qwen2-VL-based); InternVL3-8B is included only as a cross-family VLM, since no comparable InternVL-based LAM is publicly available. Our LAMs are also GUI/desktop-trained rather than physical-embodiment or game-specific. Cross-family LAM evaluation and comparison with embodied models (e.g., OpenVLA-class) are left for future work. Second limitation lies in our extension to thinking-mode foundation models is preliminary in three respects: (i) we test only a single model (Qwen3.5), and the observed asymmetry between reasoning-trace and final-answer representations may not generalize to other thinking-mode models such as DeepSeek-R1 or Qwen3 variants. (ii) the prompt and extraction setup for Qwen3.5 differ from those used for VLM/LAM extraction: Qwen3.5 was given only the Reasoning prompt, with both readouts (reasoning-span last-token and final-answer last-token) drawn from the same generation; (iii) reasoning-trace length varies across TRs, introducing residual confounds that our matched mean-pooling controls only partially. Future work will focus on systematic evaluation across multiple thinking-mode models with controlled trace-length and matched prompts. Lastly, our study compares the internal representations of the foundation models with the brain activity during naturalistic gameplay, characterizing how each system encodes the state of the game, the perceptual content and the structure relevant to the decision. However, we do not directly relate model representations to human behavioral choices, for example, button-press latencies, action selection at each time step, or trial-by-trial reaction times. Such behavioral analyses could test whether brain-aligned representations also predict moment-to-

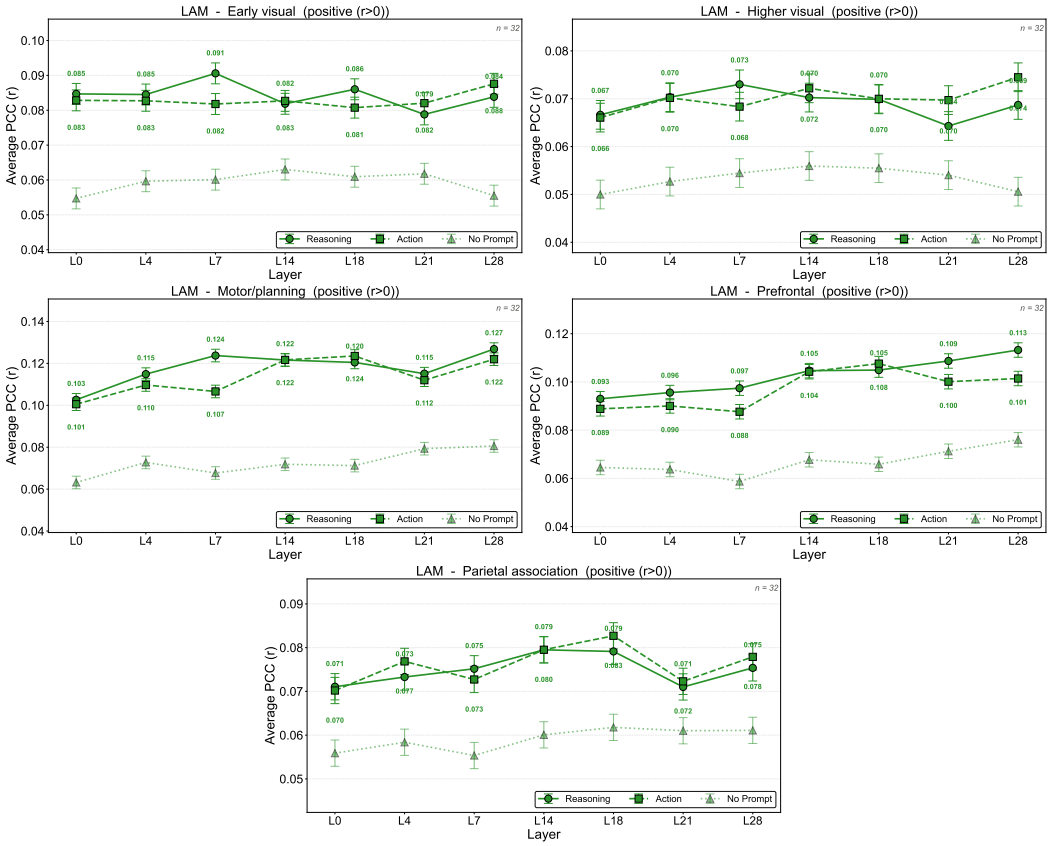


Figure 17: **Layer-wise brain alignment of LAM (UI-TARS) representations, positively predicted vertices** ($r > 0$)**.** Same as Fig. 16, restricted to voxels with $r > 0$. Error bars: mean $\pm$ SEM across $N = 32$ participants.

moment human gameplay decisions, providing a complementary axis of model–human comparison. We consider this as a natural extension to the current study.

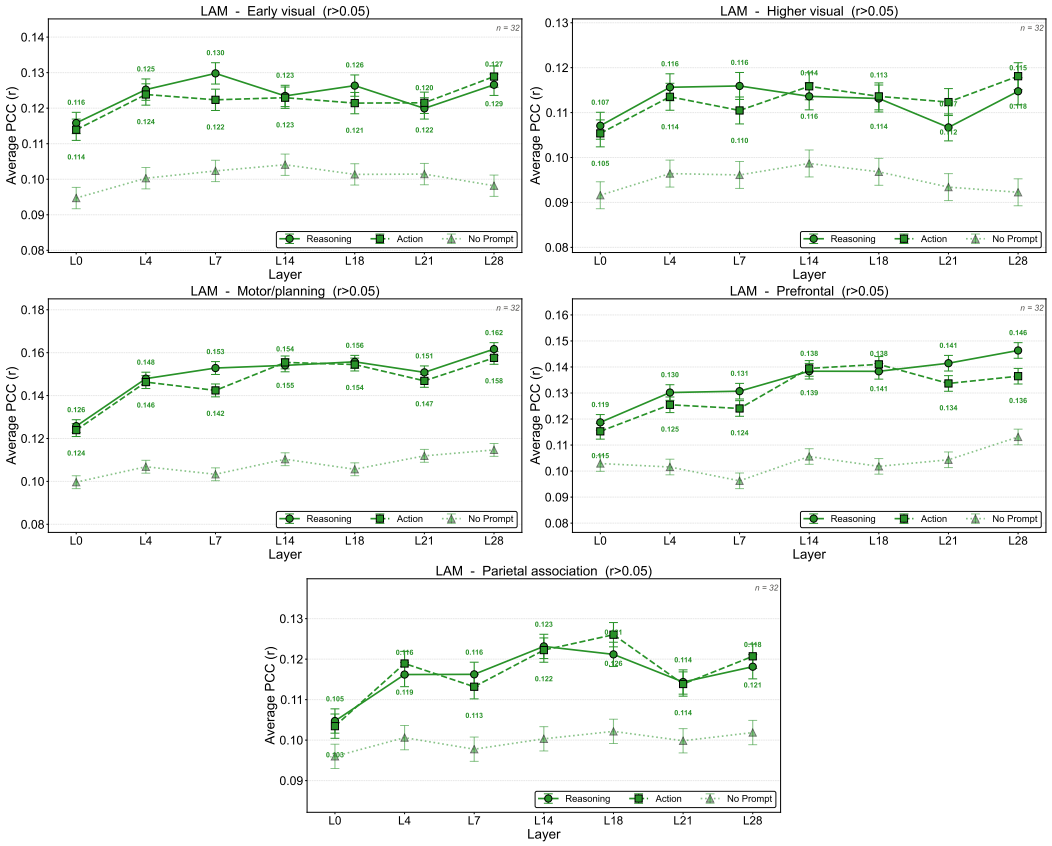


Figure 18: **Layer-wise brain alignment of LAM (UI-TARS) representations, vertices with $r > 0.05$.** Same as Fig. 16, restricted to vertices with $r > 0.05$. Error bars: mean $\pm$ SEM across $N = 32$ participants.

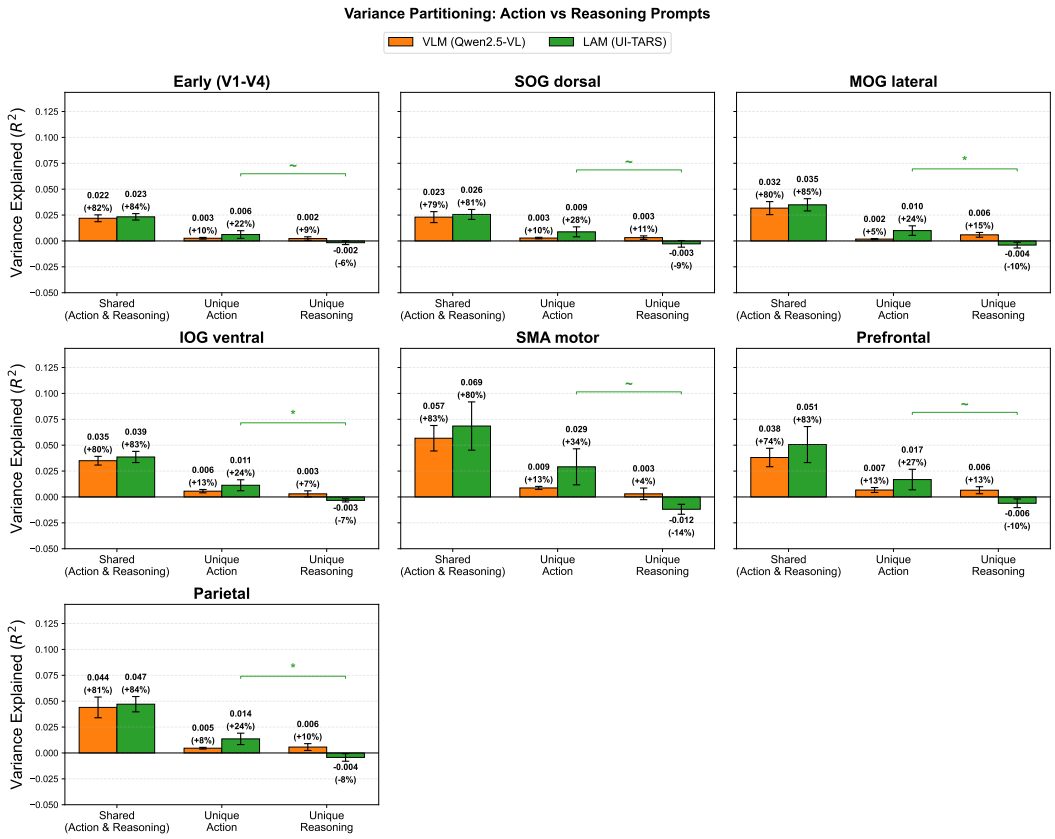


Figure 19: **Variance partitioning of action vs. reasoning prompts.** R^2 decomposed into shared variance and prompt-unique components for VLM (Qwen2.5-VL, orange) and LAM (UI-TARS-7B-DPO, green), averaged across participants. Numbers above bars show absolute R^2 and (in parentheses) percentage of joint explained variance. VLM unique variance is balanced across prompts; LAM is action-dominant, with negative unique-reasoning variance across ROIs. Brackets: VLM-vs-LAM comparison on unique reasoning, * $p < 0.05$, $\sim$ marginal ($p < 0.10$). Error bars: mean $\pm$ SEM across subjects.

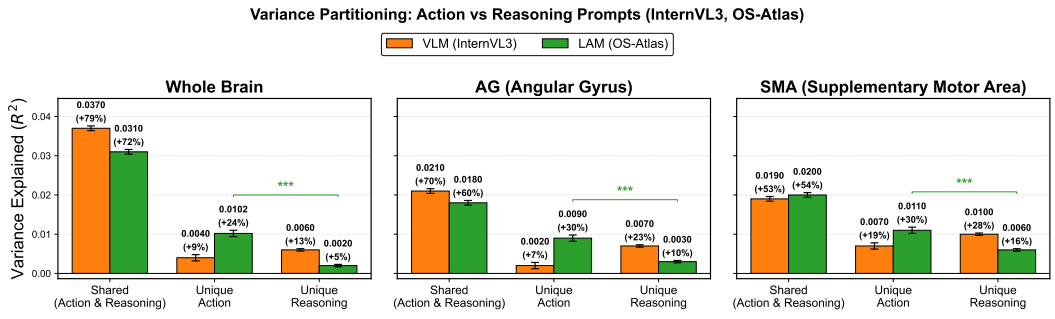


Figure 20: Brain alignment averaged across participants and voxels, using the best-performing layer for LAMs: OS-Atlas and VLMs (InternVL3). Green: LAM, Orange: VLM, * at a particular bar indicates that the model's prediction performance is significantly better than baselines.