TASK-CONDITIONED PROBING OF INSTRUCTION-TUNED MULTIMODAL LLMS: REGION-SPECIFIC BRAIN ALIGNMENT PATTERNS UNDER NATURALISTIC STIMULI

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) align better with brain activity than unimodal models. Instruction-tuned (IT) MLLMs, in which a language instruction governs how visual and auditory input is processed, produce task-specific representations. Yet their brain alignment has been evaluated only under unimodal stimuli, while studies using multimodal stimuli have relied on non-instruction-tuned models. It therefore remains unclear whether, under naturalistic multimodal input, instruction-conditioned representations capture task-related structure beyond the surface semantics of the instruction text. We address this by predicting fMRI responses recorded during naturalistic movie watching (video with audio) from MLLM representations. Across three matched pairs of pretrained and IT MLLMs given identical instructions (the pretrained models prompted via in-context learning (ICL)), IT models align better with brain activity at every layer. In all three pairs, ICL representations track instruction-text semantics ($r = 0.69$–0.78), whereas IT representations are largely decoupled from wording ($r = 0.13$–0.16) and instead exhibit significant task-conditioned organization throughout the network. This does not mean IT models ignore the instruction: they respond to which task is asked, not to how it is worded. Surprisingly, for identical movie stimuli, different instructions yield distinct regional patterns of brain predictivity: Narrative understanding in the angular gyrus and posterior cingulate cortex, spatial understanding in parietal and scene-selective regions, and emotion understanding in the middle frontal gyrus. These preferences are broadly consistent across participants. Nine video and two audio IT-MLLMs show this instruction-dependent regional structure. Together, these findings position IT-MLLMs as model organisms in which instructions derive many task-specific representations from a single stimulus, revealing which cortical regions are best predicted by which task. Because the neural data are held fixed, this provides a low-cost, model-side source of hypotheses to complement task manipulation in neuroimaging, which requires separate acquisitions per task.

1 INTRODUCTION

Brain–model correspondence under naturalistic stimuli has become a key focus (Doerig et al., 2022; Wang et al., 2023; Oota et al., 2022b; Tang et al., 2024; Oota et al., 2025a; Srijith et al., 2025), with multimodal Transformers aligning better than unimodal ones under both unimodal (Oota et al., 2022b; Popham et al., 2021) and multimodal movie stimuli (Subramaniam et al., 2024; Dong & Toneva, 2023a; Oota et al., 2025b; Sartzetaki et al., 2025). However, these models yield a single task-agnostic representation per stimulus, limiting their use as model organisms for relating task-specific representations to brain responses to the same naturalistic input (Toneva & Wehbe, 2019; Doerig et al., 2023). Instruction-tuned (IT) MLLMs, in which a language instruction guides how visual and auditory input is processed, instead generate distinct task-specific representations of the same stimulus that align with brain activity (Oota et al., 2025a; Benara et al., 2024), offering a model-side way to vary the task while neural responses remain fixed. Yet this has been tested only under unimodal stimuli, whereas multimodal-stimulus studies rely on non-IT models (Table 8, App. A).

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.

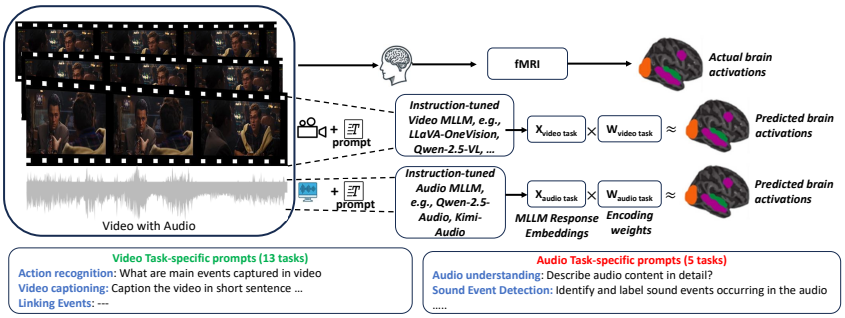


Figure 1: Leveraging instruction-tuned multimodal video and audio models for brain encoding with a diverse set of instructions. For the given movie clip, we can obtain different multimodal representations using instructions that ask the model to (i) generate the caption of the video, (ii) identify whether temporal events are present, (iii) determine the primary colors dominant in the video, etc. Using instruction-specific representations (X), we estimate the alignment using a simple linear function f (ridge regression), which maps MLLM representations to brain recordings. Here, W denotes voxelwise encoding model weights.

Alignment alone is also uninformative. Instruction-conditioned representations may merely track instruction wording rather than the task, and may improve alignment uniformly across entire brain rather than region-specifically. *This raises two questions: under naturalistic multimodal stimuli, do IT-MLLM representations reflect the task rather than the instruction text, and do they correspond to region-specific patterns of brain predictivity?*

To address this gap, we probe video and audio IT-MLLMs under naturalistic multimodal stimuli, measuring how their task-specific, language-guided representations align with brain activity from early sensory to higher-order cortical regions, and ask the following research questions: (1) Do language-guided task instructions produce task-conditioned representations in IT-MLLMs that go beyond the surface semantics of the instruction text, and how does this differ from matched pretrained (ICL) models? (2) Do IT-MLLMs show higher brain alignment than matched ICL checkpoints, and does the pattern hold across model families? (3) Under a fixed naturalistic stimulus, are different task instructions selectively predictive of distinct brain regions? (4) How are semantic categories of task instructions (e.g., narrative understanding, cognitive reasoning) associated with differences in brain alignment across language, auditory, and visual regions?

Using brain recordings from participants watching popular movies with audio (Boyle et al., 2020), we evaluate nine video IT-MLLMs, two audio IT-MLLMs, three pretrained video MLLMs with ICL, two non-IT multimodal (video+audio) models, two unimodal video models, and one unimodal audio model (Table 2), probed with 13 video and 5 audio task-specific instructions (workflow in Fig. 1).

Our key findings, corresponding to the four RQs, are: (1) Across three matched pairs, IT-MLLM representations are only weakly coupled to instruction-text semantics ($r = 0.13$–0.16) and are more consistent with task-conditioned structure, whereas ICL representations are strongly coupled ($r = 0.69$–0.78). (2a) Among IT-MLLMs, video-based models align better than audio-based ones on multimodal stimuli, a modality contrast within a fixed training paradigm. (2b) IT video MLLMs as a class show higher alignment than ICL, non-IT multimodal, and unimodal baselines. Because these families differ in scale, training data, and architecture, we treat this class-level gap as evidence that the effect is broad rather than as causal support. Attribution instead rests on three controlled within-family comparisons (InternVL3-8B, Qwen-2-VL, and Qwen-2.5-VL, each IT vs. ICL), which show positive Δ alignment at every layer while holding architecture, scale, and pretraining corpus fixed. (3) Consistent with this weak surface-semantic coupling, both video and audio IT-MLLMs produce task-conditioned representations that selectively align with distinct functional regions. E.g., audio understanding and captioning align with language areas, whereas sound event detection aligns with auditory cortex and temporal lobe. (4) Task-specific representations from IT-MLLMs thus support finer-grained mapping across cortical hierarchies; language and narrative understanding tasks, for instance, align most with angular gyrus and lateral temporal cortex, consistent with event-structure findings under naturalistic stimuli (Baldassano et al., 2017). Overall, IT-MLLMs exhibit layer- and task-dependent alignment, making them useful probes of functional specialization under naturalistic stimuli and a bridge between cognitive modeling and neuroscience. Code is in the supplementary.

2 DATASET AND MODELS

2.1 BRAIN IMAGING DATASET

We use Movie10 (Boyle et al., 2020), a naturalistic multimodal fMRI dataset from Courtois NeuroMod, comprising six participants (s1 to s6) passively watching four movies: *The Bourne supremacy (∼100 mins)*, *The wolf of wall street (∼170 mins)*, *Hidden figures (∼120 mins)* and *Life (∼50 mins)*. *Hidden Figures* and *Life* were each viewed twice, yielding repeated measurements. We use *Life* movie for testing where we average the two repetitions to reduce noise. Train and test sets are totally disjoint. fMRI volumes were acquired every 1.49 s (1 TR). Since brain-encoding models are trained separately per participant, the large number of samples per subject is more important than the number of subjects. The dataset has been used in prior multimodal brain-encoding studies (Oota et al., 2025b; d'Ascoli et al., 2026).

We use the preprocessed fsaverage6 surface data and report ROI analyses using multimodal parcellation of the human cerebral cortex based on Glasser atlas (180 ROIs per hemisphere) (Glasser et al., 2016). This includes four visual processing regions (early visual cortex (EVC), object-related areas (LOC), face-related areas (OFA) and scene-related areas (PPA)), one early auditory area (AC), and eight language-relevant regions, encompassing broader language regions: angular gyrus (AG), anterior temporal lobe (ATL), posterior temporal lobe (PTL), inferior frontal gyrus (IFG), inferior frontal gyrus orbital (IFGOrb), middle frontal gyrus (MFG), posterior cingulate cortex (PCC) and dorsal medial prefrontal cortex (dmPFC), based on the Fedorenko lab's language parcels (Milton et al., 2021; Desai et al., 2023). We show the flatmap with these labeled ROIs, in Appendix Fig. 7 and list the detailed sub-ROIs of these ROIs in App. C.

Estimating cross-subject prediction accuracy. To account for intrinsic noise in biological measurements, we estimate a cross-subject prediction accuracy following Schrimpf et al. (2021): each participant's responses are predicted from the other participants' responses using the same voxelwise encoding model (Sec. 3). Normalized brain alignment is the model's prediction accuracy divided by this cross-subject accuracy; details and ceiling maps are in App. D.

2.2 INSTRUCTION-TUNED MLLMS FOR VIDEO AND AUDIO

To test whether instruction-prompted IT-MLLMs align better with fMRI responses to multimodal stimuli, we organize the evaluation into two sets of comparisons.

Matched within-family comparisons. We compare three matched pairs of pretrained and instruction-tuned MLLMs: Qwen-2-VL (Wang et al., 2024), Qwen-2.5-VL (Qwen Team, 2025), and InternVL3 (Chen et al., 2024). Within each pair, architecture, parameter count, and pretraining corpus are identical; the checkpoints differ only in post-training. These pairs are the basis for all claims about differences associated with instruction tuning. Details are reported in Table 1.

Additional models for cross-family generalization. To test whether the matched-pair results generalize to other model families, we additionally consider six IT video MLLMs: InstructBLIPVideo (Dai et al., 2023), Video-LLaVA (Lin et al., 2024), LLaVA-NeXT-Video (Zhang et al., 2024), Qwen-2.5-Omni (Xu et al., 2025), VideoChat-R1 (Li et al., 2025), and LLaVA-OneVision (Li et al., 2025); and two IT audio MLLMs: Qwen-2.5-Audio (Chu et al., 2024) and Kimi-Audio (Kimi Team, 2024). We also experiment with two additional multimodal reference models (VILA (Lin et al., 2023) and TVLT (Tang et al., 2022)), two unimodal video models (VideoMAE (Tong et al., 2022) and TimeSFormer (Bertasius et al., 2021)), and one unimodal audio model (AST (Baade et al., 2022)). Details are reported in Table 2.

Several previous brain encoding studies have compared a wide range of language, speech, and multimodal models differing in size, architecture, training data, and objectives, with the primary goal of understanding how closely model representations align with brain-relevant semantics (Schrimpf et al., 2021; Toneva & Wehbe, 2019; Antonello et al., 2021; Oota et al., 2025b). We follow this approach for the broader cross-model comparisons, while additionally providing three matched within-family comparisons. These matched comparisons control for architecture and nominal model scale, but do not isolate the instruction-tuning objective from the accompanying post-training data.

Table 1: Matched within-family comparisons. IT denotes instruction tuning. Each pair uses the same inputs and encoding protocol; post-training data remains a coupled factor.

Model Name	IT	#Layers	Modality
Qwen-2-VL-7B	×	29	Video+Text
Qwen-2-VL-7B-IT	✓	29	Video+Text
Qwen-2.5-VL-7B	×	29	Video+Text
Qwen-2.5-VL-7B-IT	✓	29	Video+Text
InternVL3-8B	×	29	Video+Text
InternVL3-8B-IT	✓	29	Video+Text

Table 2: Additional models for cross-model generalizability. These heterogeneous comparisons do not isolate the effect of instruction tuning (IT).

Model Name	IT	#Layers	Modality
InstructBLIPVideo	✓	33	Video+Text
Video-LLaVA	✓	33	Video+Text
LLaVA-NeXT-Video	✓	33	Video+Text
VideoChat-R1	✓	29	Video+Text
LLaVA-OneVision	✓	28	Video+Text
Qwen-2.5-Omni	✓	29	Video+Audio+Text
Qwen-2.5-Audio	✓	29	Audio+Text
Kimi-Audio	✓	29	Audio+Text
VILA	×	29	Video+Audio
TVLT	×	12	Video+Audio
VideoMAE	×	24	Video
TimeSFormer	×	12	Video
AST	×	24	Audio

Table 3: Instructions for multimodal audio tasks.

Task	Description
Audio Understanding	Can you describe the audio content in detail?
Audio Comprehension	What are people doing in the audio?
Audio Captioning	Caption the audio in a short sentence.
Sound Event Detection	Identify and label the sound events occurring in the audio.
Speaker Identification	Who is speaking in the audio?

Table 4: Instructions for various multimodal video tasks.

Task	Description
Action Recognition	What are the main events captured in the video?
Video Understanding	Can you describe the video content in detail?
Visual Question Answering	How many people are in the video, and what are they doing?
Video Captioning	Caption the video in a short sentence.
Object and Scene Recognition	What are the main objects and people visible in the video? Describe each one briefly.
Commonsense Reasoning	Why did the character take this action? What could have motivated them to do this?
Spatial Understanding	Where is this video taken from? What place/landmark is shown in the video?
Temporal Ordering	Step-by-step describe the activity shown in the video.
Video reasoning	What is unusual about this video?
Narrative Understanding	Summarize the main storyline of the movie. What is the central conflict, and how is it resolved?
Emotion and Sentiment Analysis	What emotions do the characters express during the video? How does the video make you feel overall?
Global Appearance	Describe changes in characters' appearances throughout the video, including any noticeable outfit changes.
Linking Events	Explain how an early event in the video influences later developments.

2.3 Natural Language Instructions and Feature Extraction from IT-MLLMs

Video-specific tasks. We use 13 task instructions (Table 4) inspired by VideoInstruct100K (Maaz et al., 2024), chosen for diversity and applicability to any video regardless of frame content. Sample generated outputs for all video MLLMs appear in Tables 21–26 (App. V).

We extract instruction-specific representations by feeding a video and a task instruction and taking hidden states for the language-guided response. Each matched pretrained checkpoint receives the identical video and instructions; we call this the in-context-learning (ICL) setting, since the task is specified through the prompt alone, without instruction tuning or in-context demonstrations. For all IT and ICL video MLLMs, we average per-layer hidden states over generated output tokens, yielding one embedding per layer per video–instruction pair. Non-IT multimodal models (TVLT, VILA) receive video and audio; unimodal video models (VideoMAE, TimeSFormer) receive video only. All models are used zero-shot, without fine-tuning on fMRI data.

Audio-specific tasks. We use five natural instructions for audio input to audio IT-MLLMs (Table 3), with sample generated outputs in Tables 27 and 28 (App. V). Representations are extracted as for video. We input an audio clip with a task instruction and take the embeddings for the language-guided response. AST receives audio only.

3 Methodology

Voxel-wise encoding model. We train banded ridge regression based voxel-wise encoding models (la Tour et al., 2022) to predict the fMRI brain activity associated with the stimulus representations obtained from task-specific instructions from instruction-tuned and other baseline models. Banded ridge regression optimizes a different regularization hyperparameter per feature space, and decomposes the explained variance over feature spaces. This decomposition helps in identifying which task-specific instruction contributes most to the explainable variance in different brain regions. We employ z-score thresholding separately for both input stimulus representations and brain recordings for training and test datasets. For each subject, we account for the delay in the hemodynamic response by modeling hemodynamic response function using a finite impulse response (FIR)

Table 5: IT vs. ICL models: representational similarity with instruction-text semantics for three matched pairs. *** p<0.001, FDR-significant across layers; ns = not significant after FDR correction. Magnitude is the ICL/IT ratio at L25.

Metric	Model	Layer 5	Layer 14	Layer 25	Winner	Magnitude
	IT (Qwen-2.5-VL-Instruct)	Pearson $r = 0.077$ (ns)	0.130 (ns)	0.143 (ns)	ICL	Weak
	ICL (Qwen-2.5-VL)	0.690***	0.680***	0.777***		$\sim$5.4$\times$ stronger
Semantic correlation	IT (Qwen-2-VL-Instruct)	Pearson $r = 0.091$ (ns)	0.116 (ns)	0.158 (ns)	ICL	Weak
	ICL (Qwen-2-VL)	0.548***	0.626***	0.693***		$\sim$4.4$\times$ stronger
Semantic correlation	IT (InternVL3-8B-Instruct)	Pearson $r = 0.052$ (ns)	0.101 (ns)	0.126 (ns)	ICL	Weak
	ICL (InternVL3-8B)	0.614***	0.587***	0.708***		$\sim$5.6$\times$ stronger

Table 6: Within-video task-label permutation test. Δ_{task}: mean within-task minus between-task representation similarity (range across 29 layers); q: BH-corrected across layers; depth trend: Spearman correlation between layer index and Δ_{task}.

Matched Pair	Model	Δ_{task}	q	Depth ρ	p
Qwen-2.5-VL-7B	IT	0.232–0.472	< 0.001	0.91	< 0.001
	ICL	0.018–0.081	0.087–0.412	0.19	0.323
Qwen-2-VL-7B	IT	0.186–0.391	< 0.001–0.004	0.84	< 0.001
	ICL	0.011–0.067	0.104–0.536	0.14	0.468
InternVL3-8B	IT	0.204–0.428	< 0.001–0.002	0.88	< 0.001
	ICL	0.016–0.074	0.092–0.487	0.21	0.274

per voxel with 5 temporal delays corresponding to $\sim$7.5 seconds (Huth et al., 2022). Formally, at each time step t, we encode the stimuli as $X_t \in \mathbb{R}^D$ and brain region voxels $Y_t \in \mathbb{R}^V$, where D denotes the dimension of the concatenation of delayed 5 TRs, and V denotes the number of voxels. Overall, with N such TRs, we obtain N training examples. Detailed hyper-parameter settings are in App. E.

Evaluation metrics. We evaluate our models using Pearson Correlation (PC), which is a standard metric for evaluating brain alignment (Jain & Huth, 2018; Schrimpf et al., 2021; Goldstein et al., 2022). Let TR be number of TRs in the test set. Let $Y = \{Y_i\}_{i=1}^{TR}$ and $\hat{Y} = \{\hat{Y}_i\}_{i=1}^{TR}$ denote actual and predicted value vectors for a single voxel. Thus, Y and $\hat{Y} \in \mathbb{R}^{TR}$. PC is computed as correlation between the model's predictions $\hat{Y}$ and neural recordings Y. To quantify the model predictions, the resulting model prediction correlations are divided by the estimated cross-subject prediction accuracy; resulting in a standardized measure of performance referred to as normalized brain alignment. To calculate *normalized alignment*, we select voxels with cross-subject prediction accuracy ≥ 0.05, in line with previous works (Popham et al., 2021; la Tour et al., 2022).

4 Results

[RQ1]: ICL-MLLM representations strongly correlate with instruction-text semantics; IT-MLLM representations show weak semantic coupling and are more task-conditioned.

To identify representational differences between IT and ICL MLLMs, we first compute a 13$\times$13 semantic-similarity matrix over the video task instructions using two independent text embedding models, all-MiniLM-L6-v2 (Reimers & Gurevych, 2019) and MPNet (Song et al., 2020) (App. G, Fig. 9). Under MiniLM, similarity ranges from 0.15 to 0.85 (mean 0.42 ± 0.18). High-similarity pairs (> 0.60) include Action Recognition–Video Understanding (0.68), Object & Scene Recognition–Action Recognition (0.68), and Video Understanding–Visual Question Answering (0.65), whereas Commonsense Reasoning is dissimilar to most tasks and Spatial Understanding is dissimilar to Emotional and Narrative tasks. This confirms our task-instruction set contains a broad range of semantic relatedness across task instructions.

ICL vs. IT MLLM representations. For the 13 video tasks, we extracted language hidden states from three matched pairs across all 29 layers for a pretrained (ICL) checkpoint and its IT variant, using identical video inputs with different instructions to isolate instruction effects. We show detailed results for representative early, middle, and late layers (L5, L14, L25) in App. G Fig. 10. While we present detailed results for three representative layers, we computed correlations and clustering metrics across all 29 layers to ensure robustness.

Semantic coupling differs sharply between ICL and IT. Table 5 summarizes semantic coupling and functional-vs-semantic clustering differences for representative layers in the three matched

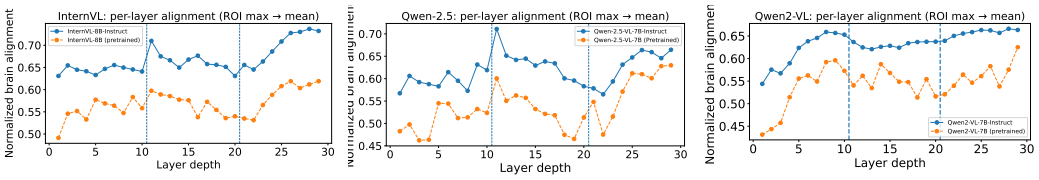


Figure 2: Normalized brain alignment before vs. after instruction tuning.

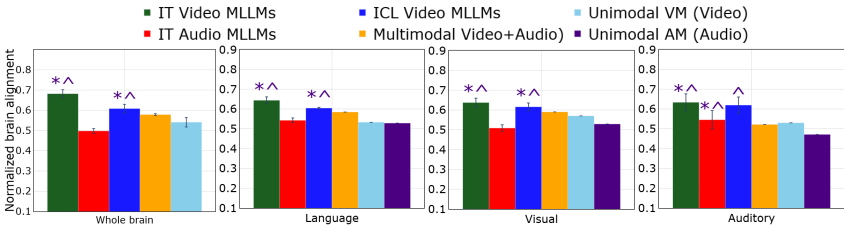


Figure 3: Average normalized brain alignment of IT video MLLMs vs IT audio MLLMs vs ICL video MLLMs vs multimodal and unimodal models across whole brain, language, visual and auditory regions. Error bars indicate the standard error of the mean across participants. $*$ implies that IT MLLM embeddings are significantly better than multimodal models and $\wedge$ means that IT MLLM embeddings are significantly better unimodal models with p≤ 0.05.

pairs. For the 13 tasks, we perform representational similarity analysis (RSA) by computing correlations between the upper triangles of the 13×13 semantic-similarity matrix (from App. G Fig. 9) and the corresponding representation-similarity matrix derived from MLLM hidden states (same videos, same pipeline), at each layer for both the ICL and IT checkpoint of every pair. We find that in all three pairs, IT representations are only weakly coupled to instruction semantics ($r = 0.13$–0.16 at L25), and no layer remains significant after FDR correction (App. G Table 10). In contrast, ICL representations show strong coupling ($r = 0.69$–0.78 at L25; FDR $q < 0.001$), 4.4–5.6$\times$ stronger than their IT counterparts, indicating substantially greater sensitivity to instruction-text semantics (ARI ≈ -0.12; App. Table 11).

IT representations exhibit significant task-conditioned structure across layers. Because weak semantic coupling alone does not establish task-conditioned organization, we tested this directly with a within-video task-label permutation analysis. At each layer we defined Δ_{task} as the difference between mean within-task and between-task representation similarity, generating a stimulus-preserving null by permuting task labels within each video 10,000 times. In all three matched pairs (Table 6), Δ_{task} is positive and significant at every layer for the IT model (Qwen-2.5-VL: 0.232–0.472; Qwen-2-VL: 0.186–0.391; InternVL3: 0.204–0.428; all BH-corrected $q \leq 0.004$) but small and non-significant for the ICL model (0.011–0.081; $q \geq 0.087$), indicating task-conditioned structure beyond variation attributable to video identity. It also increases with depth in every IT model (Spearman $\rho = 0.84$–0.91, all $p < 0.001$), with no such trend in ICL models ($\rho = 0.14$–0.21, $p \geq 0.27$). With the RSA and clustering results, this indicates that ICL representations are organized largely by instruction-text semantics, whereas IT representations carry task-conditioned structure throughout the network (Zhu et al., 2025). See App. G and H.

[RQ2]: IT MLLMs SHOW HIGHER BRAIN ALIGNMENT THAN THEIR MATCHED ICL COUNTERPARTS, AND THE PATTERN HOLDS ACROSS MODEL FAMILIES.

We measure normalized brain alignment at the whole-brain level and within language, visual, and auditory ROIs on the Movie10 dataset. We first compare each IT model with its matched ICL checkpoint, and then assess whether the pattern holds across a broader, uncontrolled set of models.

Matched pairs. **Effect of instruction tuning**: We compare three same-family pairs: Qwen-2.5-VL-7B, Qwen-2-VL-7B, and InternVL3-8B (ICL vs. Instruct), using identical video inputs and the same prompt ("Describe the video"), and compute per-layer Δ alignment as normalized brain alignment (IT) $-$ normalized brain alignment (ICL). In all three pairs, Δ is positive at every layer (Fig. 2; Table 16), with mean gains of 0.153 (Qwen-2.5-VL), 0.086 (Qwen-2-VL), and 0.100 (InternVL3). The depth profile of the gain differs by family: for Qwen-2.5-VL it is largest in early layers and decreases with depth ($\Delta = 0.207$ at L1 to 0.088 at L29; Spearman $\rho = -0.77$, $p = 1.2 \times 10^{-6}$), for InternVL3 it is broadly distributed across layers ($\Delta = 0.056$–0.140; $\rho = 0.34$, $p = 0.07$), and

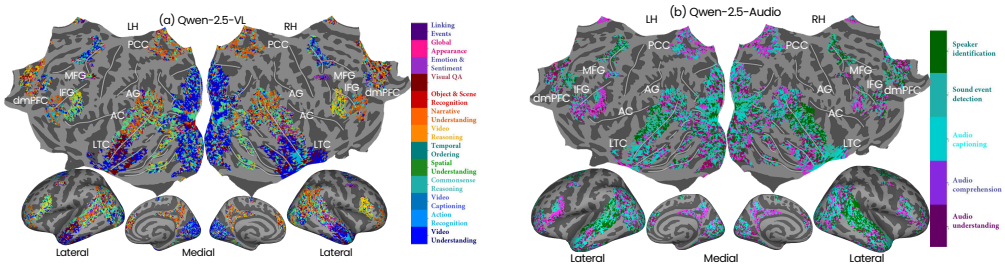


Figure 4: Each voxel is color-coded with the instruction that led to the highest normalized brain alignment. The voxels are projected onto the flattened cortical surface of the 'fsaverage' subject. (Left): video MLLM (Qwen-2.5-VL). (Right): audio MLLM (Qwen-2.5-Audio).

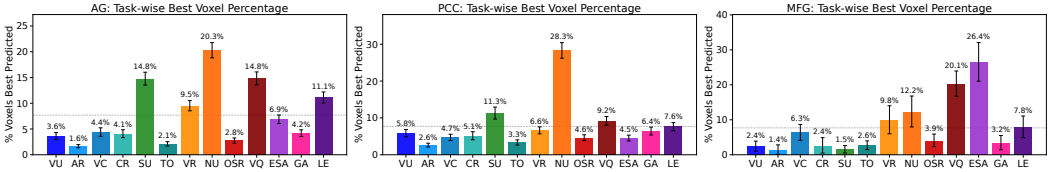


Figure 5: % of best predicted voxels per task for an IT video MLLM across 3 lang ROIs. Error bars denote 95% confidence intervals from 10,000 within-ROI voxel-bootstrap resamples (voxels pooled across six participants). The dotted line marks the equal-preference reference (7.69%). Plots for remaining language, visual and auditory ROIs are shown in Appendix Fig. 15

for Qwen-2-VL the gain is also broadly distributed across layers ($\Delta = 0.038$–0.132; $\rho = 0.11$, $p = 0.566$). Instruction tuning is thus associated with higher alignment throughout the network, while the layers that benefit most are family-specific.

No-prompt baseline. Task-specific prompting boosts alignment over a no-prompt baseline (0.746 vs 0.639, +16.7% for Qwen-2.5-VL), showing gains beyond generic video representations.

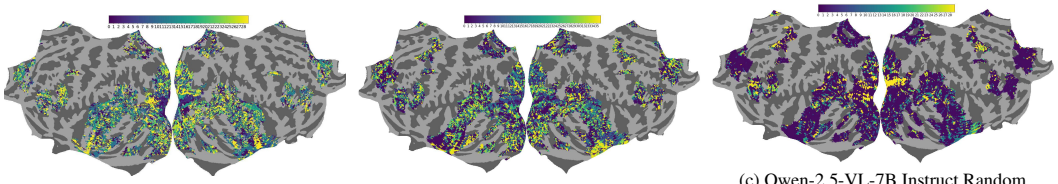
Generalization across model families. We next compare the broader set of IT video and audio MLLMs with ICL video MLLMs, non-instruction-tuned multimodal models (VILA, TVLT), and unimodal baselines (VideoMAE, TimeSFormer, AST), averaging normalized alignment across instructions (13 video, 5 audio), participants, and the best layer (Fig. 3). Because these models differ in architecture, scale, and training data, we treat these comparisons as descriptive. At the whole-brain level, IT video MLLMs show higher alignment than ICL models ($\sim$9%) and than non-instruction-tuned multimodal and unimodal baselines ($\sim$15%) (paired t-tests, FDR-corrected). The same pattern holds in language and visual regions. IT audio MLLMs show lower whole-brain alignment than the video-based models, with gains relative to baselines confined mainly to auditory cortex and MFG. Results for individual models and sub-regions are in App. I.

Whole brain analysis. Fig. 3 (a) provides these insights: (i) At the whole-brain level, the paired t-test reveals that the differences in brain alignment between IT video MLLMs and ICL models, the non-instruction-tuned multimodal and unimodal models are statistically significant. In particular, all IT video MLLMs achieve over $\sim$9% improvement in brain alignment compared to ICL models, and $\sim$15% improvement compared to other baselines. This contrasts with prior findings on IT image-based MLLMs, which demonstrated comparable performance to multimodal models when evaluated on unimodal image stimuli (Oota et al., 2025a), suggesting that IT video MLLMs are more effective at capturing brain-relevant representations. (ii) IT audio MLLM embeddings show less alignment compared to non instruction-tuned multimodal and unimodal video models. These findings indicate that IT video MLLM models capture brain-relevant representations and contain additional information beyond the ICL, non-instruction-tuned multimodal and unimodal models.

Language, visual and auditory region analysis. We also present the average normalized brain alignment across language, visual and auditory regions in Fig. 3 (b, c & d). The results from the paired t-tests across participants are consistent with whole-brain performance both in the language and visual regions, i.e., IT video MLLMs embeddings exhibit significantly higher alignment in both language and visual regions compared to ICL video MLLMs, and all baselines. On the other hand, IT audio MLLM embeddings show significant alignment primarily in the auditory cortex and the middle frontal gyrus; when compared to non-instruction-tuned multimodal and unimodal models. Results for individual models are shown in Fig. 12, detailed language, visual and auditory

Table 7: Modality control: Mean ± SD normalized alignment across subjects using Qwen-2.5-Omni.

Modality	Whole Brain	Language	Visual	Auditory
Video + Audio	**0.670 ± 0.021**	**0.685 ± 0.013**	**0.689 ± 0.024**	**0.681 ± 0.019**
Video only	0.601 ± 0.011	0.615 ± 0.027	0.629 ± 0.027	0.629 ± 0.027
Audio only	0.534 ± 0.030	0.545 ± 0.019	0.521 ± 0.025	0.560 ± 0.011



(a) Qwen-2.5-VL-7B Instruct (b) Qwen-2.5-VL-3B Instruct (c) Qwen-2.5-VL-7B Instruct Random Prompt

Figure 6: Layer-wise alignment: Each voxel is color coded with the MLLM layer number (out of 29) that led to the highest normalized brain alignment. The voxels are projected onto the flattened cortical surface of average across subjects on 'fsaverage' surface.

sub-regions are shown in Fig. 13 and 14 in App. I. These results suggest that IT video MLLMs more effectively capture brain-relevant multimodal representations, particularly when processing multimodal stimuli.

[RQ3]: VIDEO AND AUDIO IT-MLLMS SUCCESSFULLY DIFFERENTIATE TASK-SPECIFIC INSTRUCTIONS.

To examine which instructions best predict brain activity in each region, we assign each voxel the instruction with the highest normalized brain alignment. Because participants viewed the movies passively, these analyses characterize which model-side instructions best predict each region's responses; they do not imply that participants performed the corresponding tasks.

Instruction-tuned video MLLMs. Fig. 4 (left) shows, for Qwen-2.5-VL, the instruction that best predicts each voxel, averaged across participants. We make the following observations: (i) Narrative Understanding best predicts the largest share of voxels in higher-order regions, including AG (20.3%), PCC (28.3%), and dmPFC (22.7%). (ii) Spatial Understanding and Visual Question Answering best predict voxels in language-related regions such as AG and the posterior temporal lobe, where Spatial Understanding is the top instruction (18.9%); in ATL, Visual Question Answering (19.5%) and Video Understanding (20.5%) predominate. (iii) Video Understanding best predicts early visual cortex (20.1%), a preference reproduced in all six participants, whereas Emotion and Sentiment Analysis best predicts MFG (26.4%). Percentages are shares of ROI voxels (chance = 7.7%; App. Table 13); participant-level consistency is reported in App. Table 14.

ROI-differentiated task-preference patterns. Fig. 5 presents three illustrative higher-order ROIs, while the complete ROI × task-preference distributions and 95% bootstrap confidence intervals are reported in Appendix Table 13. From Fig. 5, we observe distinct task-region preferences: (i) AG is dominated by Narrative Understanding (NU, 20.3%), Visual QA (VQ, 14.8%), and Spatial Understanding (SU, 14.7%), consistent with its role in narrative integration and cross-modal semantic reasoning. (ii) PCC shows the strongest dominance for Narrative Understanding (NU, 28.3%) followed by Spatial Understanding (SU, 11.3%) and Visual QA (VQ, 9.2%), consistent with its role in episodic memory and discourse-level processing. (iii) MFG is dominated by Emotion/Sentiment Analysis (ESA, 26.4%), Visual QA (VQ, 20.1%), and Narrative Understanding (NU, 12.2%), consistent with its role in higher-order reasoning, emotional integration, and cognitive control. These results suggest that IT video MLLMs more effectively capture brain-relevant multimodal representations, and that different task instructions preferentially engage distinct higher-order ROIs in ways consistent with known functional neuroanatomy. Plots for remaining ROIs are shown in Appendix Fig. 15. These findings suggest that IT video MLLMs not only capture modality-specific representations (e.g., visual, linguistic), but also encode task-specific instructions involving semantic integration and event structure (like video understanding). We observe similar performance gains in other IT video MLLMs; flatmaps showing task-specific encoding performance for average of subjects are shown in Figs. 19 and 20 in App. K.

Instruction-tuned audio MLLMs. Fig. 4 (right) shows a clear distinction between different audio tasks. Speaker identification and sound detection are aligned with the auditory cortex (AC), while audio understanding activates higher-level regions like the inferior temporal (IT) cortex and inferior

frontal gyrus (IFG). In contrast, speaker identification shows very sparse and scattered alignment, with some unexpected activation in the primary visual cortex (V1), suggesting it does not strongly reflect brain-relevant semantic processing. Fig. 21 in App. K shows similar brainmap for Kimi-Audio. Plots for ROIs associated with task are shown in Appendix Fig. 16.

Modality control. Using Qwen-2.5-Omni for a controlled comparison, video+audio consistently outperforms unimodal inputs, video exceeds audio in most regions, and audio performs best in auditory cortex. Overall this shows that modality drives alignment beyond architecture and scale.

IT-MLLMs capture layer-wise cortical hierarchy. In Fig. 6, for each voxel, we assign the layer with highest normalized brain alignment and visualize this on cortical maps. For Qwen-2.5-VL-7B (Fig. 6, left), we observe a clear gradient: (i) early sensory areas (both visual and auditory) align with lower layers, capturing low-level features; (ii) high-level visual (LOC and PPA) and lang regions (superior temporal sulcus (STS) and AG) align with middle-deep layers, reflecting increasing abstraction and semantically rich representations; (iii) lang regions (IFG, ATL, AG) align most with deepest layers. Overall, IT-MLLMs show a layered structure aligned with brain's hierarchy. See Fig. 22 in App. L for other models.

Further, we validate this hierarchy under two settings: (i) scaling the model (Qwen-2.5-VL 3B vs. 7B Instruct) and (ii) a non-language prompt control. Fig. 6 (middle) shows the 3B-Instruct model, and (right) the 7B model with a random prompt (like ##### #### ####). We observe: (i) Non-language control: no hierarchical gradient, with early layers dominating across cortex; (ii) 3B-Instruct: the same early-to-mid/late cortical gradient as the larger model.

[RQ4]: SEMANTIC TASK GROUPS SHOW DISTINCT BRAIN-ALIGNMENT PATTERNS IN IT VIDEO MLLMS.

We group 13 video tasks into 5 cognitive categories (perceptual, cognitive reasoning and integration, spatiotemporal understanding, language/narrative understanding, social/affective understanding) and observe clear regional distinctions (Appendix O Fig. 28).

Language/narrative tasks show broad activation in temporal–parietal regions (e.g., bilateral temporal lobes, AG, posterior STS, PCC), consistent with multimodal integration needed for narrative comprehension. This aligns with (Mar, 2011; Baldassano et al., 2017). **Cognitive reasoning** splits by type: commonsense engages temporal cortex, AG, and higher-order visual areas for semantic processing and world knowledge; video reasoning aligns with early visual regions (V1, V2, V3) for visual perception and motion processing; linking-events tasks involve early auditory cortex, ATL, IFG, and AG for integrating temporal, linguistic, and episodic information. Overall, these category-level patterns indicate that different reasoning tasks are associated with distinct region-wise alignment profiles across modalities and timescales.

Similar category-specific patterns appear for other groups (App. O), indicating task-dependent IT-MLLM representations with distinct regional alignment profiles. We also quantify overlap and dissociation in variance explained by different task instructions via variance partitioning (App. P).

5 CONCLUSION

Using task-conditioned video/audio IT-MLLM representations, we predict fMRI activity during naturalistic viewing and compare alignment across modalities and ROIs. IT video MLLMs outperform audio, ICL, unimodal, and non-instruction-tuned baselines. Key findings: (1) ICL representations strongly couple to instruction wording, while IT representations are less sensitive to surface semantics. (2) IT video MLLMs show strong whole-brain alignment; IT audio MLLMs gains are limited to auditory cortex and a language region (MFG), with performance near non-instruction baselines. This modality asymmetry is consistent with prior studies of audio–brain alignment patterns (Oota et al., 2024a; 2025b). (3) Task-specific instructions yield distinct task-conditioned representations in both video and audio MLLMs, producing region-differentiated alignment patterns not observed under unimodal stimuli in prior work (Oota et al., 2025a). This supports using instructions as a controlled lever to vary task goals while holding the stimulus fixed. (4) Layer-wise alignment mirrors brain hierarchy, with shallow layers matching sensory areas and deeper layers matching higher-order regions. See App. Q for extended discussion, and App. U for limitations.

6 AI USAGE DISCLOSURE

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

This work uses the publicly released Movie10 dataset from the Courtois NeuroMod databank (Boyle et al., 2020). All fMRI data were collected by the original data providers under their institutional ethics approval with informed participant consent; we performed no new data collection with human subjects and accessed only de-identified recordings under the dataset's terms of use. Participants are identified only by anonymous codes (s1–s6), and no attempt was made to re-identify individuals. The movie stimuli are copyrighted commercial films; we neither redistribute them nor release any derived stimulus content. All MLLMs evaluated are publicly available checkpoints used in accordance with their licenses. Our analyses are correlational, characterizing which model representations predict recorded brain responses; they do not decode private mental content, and the encoding models are not intended for any diagnostic or individual-level inference.

REPRODUCIBILITY STATEMENT

Both the naturalistic stimuli (movies) and the fMRI recordings used in this study are publicly available, with preprocessing steps and experimental settings described in Section 2.1 and further detailed in Appendix C. Task-specific instruction representations from instruction-tuned video and audio MLLMs, as well as in-context learning video MLLMs, are described in Section 2.3. Implementation details of voxelwise brain encoding models and evaluation metrics are provided in Section 3, with hyperparameters listed in Appendix E. To facilitate reproducibility, we release anonymized source code for all models (instruction-tuned, in-context learning, multimodal, unimodal), brain encoding, and evaluation in the supplementary zip file.

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.

Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36, 2024.

Khai Loong Aw and Mariya Toneva. Training language models to summarize narratives improves brain alignment. In *The Eleventh International Conference on Learning Representations*, 2023.

Alan Baade, Puyuan Peng, and David Harwath. Mae-ast: Masked autoencoding audio spectrogram transformer. *Interspeech 2022*, 2022.

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.

Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W Pillow, Uri Hasson, and Kenneth A Norman. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721, 2017.

Pascal Belin, Robert J Zatorre, Philippe Lafaille, Pierre Ahad, and Bruce Pike. Voice-selective areas in human auditory cortex. *Nature*, 403(6767):309–312, 2000.

Vinamra Benara, Chandan Singh, John X Morris, Richard Antonello, Ion Stoica, Alexander G Huth, and Jianfeng Gao. Crafting interpretable embeddings by asking llms questions. *Advances in Neural Information Processing Systems*, 36, 2024.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.

Julie A. Boyle, Basile Pinsard, Amal Boukhdhir, Sylvie Belleville, Simona Brambatti, Jeni Chen, Julien Cohen-Adad, André Cyr, Adrian Fuente, Pierre Rainville, and Pierre Bellec. The courtois project on neuronal modelling - first data release. In *26th OHBM annual meeting*. Organization for Human Brain Mapping (OHBM), 2020. URL `https://publications.polymtl.ca/50613/`.

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.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

William Jay Conover. *Practical nonparametric statistics*, volume 350. john wiley & sons, 1999.

Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, 15(1):9383, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 2023.

Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Tribe: Trimodal brain encoder for whole-brain fmri response prediction. In *International Conference on Learning Representations*, volume 2026, pp. 152742–152759, 2026.

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.

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.

Adrien Doerig, Tim C Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737*, 2022.

Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lindsay, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24(7):431–450, 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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Changjiang Gao, Zhengwu Ma, Jiajun Chen, Ping Li, Shujian Huang, and Jixing Li. Increasing alignment of large language models with language processing in the human brain. *Nature computational science*, 5(11):1080–1090, 2025.

Christopher R Genovese. A bayesian time-course model for functional magnetic resonance imaging data. *Journal of the American Statistical Association*, 95(451):691–703, 2000.

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.

L Hubert and P Arabie. Comparing partitions journal of classification 2 193–218. *Google Scholar*, 193, 1985.

Alexander G Huth, Shinji Nishimoto, An T Vu, and T Dupre La Tour. Gallant lab natural short clips 3t fmri data. *G-Node doi*, 10, 2022.

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. *Advances in Neural Information Processing Systems*, 31, 2018.

Team Kimi Team. Kimi-audio technical report, 2024.

Tom Dupré la Tour, Michael Eickenberg, Anwar O Nunez-Elizalde, and Jack L Gallant. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728, 2022.

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.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns llms to the human brain. *First Conference on Language Modeling*, 2024.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024.

Raymond A Mar. The neural bases of social cognition and story comprehension. *Annual Review of Psychology*, 62(1):103–134, 2011.

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.

Omer Moussa, Dietrich Klakow, and Mariya Toneva. Improving semantic understanding in speech language models via brain-tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KL8Sm4xRn7.

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, Agarwal Veeral, Marreddy Mounika, Gupta Manish, and Raju Surampudi Bapi. Speech taskonomy: Which speech tasks are the most predictive of fmri brain activity? In *24th INTERSPEECH Conference*, 2023.

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.

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.

Sara F Popham, Alexander G Huth, Natalia Y Bilenko, Fatma Deniz, James S Gao, Anwar O Nunez-Elizalde, and Jack L Gallant. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, 24(11):1628–1636, 2021.

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

Aniketh Janardhan Reddy and Leila Wehbe. Can fmri reveal the representation of syntactic structure in the brain? *Advances in Neural Information Processing Systems*, 34:9843–9856, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Christina Sartzetaki, Gemma Roig, Cees GM Snoek, and Iris IA Groen. One hundred neural networks and brains watching videos: Lessons from alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/pdf?id=LM4PYXBId5.

Oliver Schoppe, Nicol S Harper, Ben DB Willmore, Andrew J King, and Jan WH Schnupp. Measuring the performance of neural models. *Frontiers in computational neuroscience*, 10:10, 2016.

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.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Padakanti Srijith, Khushbu Pahwa, Radhika Mamidi, Bapi Raju Surampudi, Manish Gupta, and SUBBA REDDY OOTA. Aligning text/speech representations from multimodal models with MEG brain activity during listening. In *The 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://openreview.net/forum?id=1I1z0FlEZk.

V Subramaniam, C Wang, A Barbu, G Kreiman, and B Katz. Revealing vision-language integration in the brain with multimodal networks. In *International Conference on Machine Learning*. International Conference on Machine Learning (ICML), 2024.

Jingyuan Sun and Marie-Francine Moens. Fine-tuned vs. prompt-tuned supervised representations: which better account for brain language representations? In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 5197–5205, 2023.

Jingyuan Sun, Xiaohan Zhang, and Marie-Francine Moens. Tuning in to neural encoding: Linking human brain and artificial supervised representations of language. In *ECAI 2023*, pp. 2258–2265. IOS Press, 2023.

Jerry Tang, Meng Du, Vy Vo, Vasudev Lal, and Alexander Huth. Brain encoding models based on multimodal transformers can transfer across language and vision. *Advances in Neural Information Processing Systems*, 36, 2024.

Zineng Tang, Jaemin Cho, Yixin Nie, and Mohit Bansal. Tvlt: Textless vision-language transformer. *Advances in Neural Information Processing Systems*, 35:9617–9632, 2022.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 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.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 35:10078–10093, 2022.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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 R 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.

Nishitha Vattikonda, Aditya R Vaidya, Richard J Antonello, and Alexander G Huth. Brainwavlm: Fine-tuning speech representations with brain responses to language. *arXiv preprint arXiv:2502.08866*, 2025.

Aria Wang, Michael Tarr, and Leila Wehbe. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. *Advances in Neural Information Processing Systems*, 32:15501–15511, 2019.

Aria Y Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11445–11465, 2023.

Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. Event perception: a mind-brain perspective. *Psychological Bulletin*, 133(2):273, 2007.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Overview of Appendix Sections

- App. A: Overview of multimodal model evaluation settings in brain encoding studies
- App. B: Related work
- App. C: Detailed sub-ROIs of language, visual and auditory regions
- App. D: Cross-subject prediction accuracy
- App. E: Implementation details for reproducibility
- App. F: Statistical Significance
- App. G: Correlation Between Instruction Semantics and Model Representations
- App. H: Representational Differences Between IT and ICL Models
- App. I: Effectiveness of instruction-tuned video MLLMs vs audio MLLMs vs multimodal vs unimodal representations for various brain regions
- App. J: Contrasting Instruction-tuned Video MLLMs with In-context Learning video MLLMs
- App. K: Brain Maps for Task-specific instructions
- App. L: Brain Maps showing Layer-wise Details for Video Instruction-based MLLMs
- App. M: Validation of "hierarchical correspondence" between Model Layers and Brain Regions
- App. N: Self-Controlled Experiments: Comparing Models Before and After Instruction Tuning
- App. O: Details of Semantic Task Group Analysis
- App. P: Details of Explained Variance Partitioning
- App. Q: Extended Discussion
- App. R: Discussion on Controlling Architectural and Pretraining Differences Across MLLMs

- App. S: Normalized brain alignment: cross-subject vs. repeat-based EV ceiling.
- App. T: Cross Subject Generalization
- App. U: Limitations
- App. V: Model generated outputs across instructions

A Overview of multimodal model evaluation settings in brain encoding studies

Recent studies on multimodal models in brain encoding commonly consider two settings, as shown in Table 8: (i) multimodal models evaluated with unimodal stimuli, and (ii) multimodal models evaluated with multimodal stimuli. In the former setting, brain recordings are obtained from unimodal image stimuli, but representations from multimodal models, which integrate modalities such as vision and language, achieve a higher degree of brain alignment compared to vision-only models. In the latter setting, where brain recordings are obtained from multimodal stimuli (video with audio), studies show that multimodal models exhibit higher degree of brain alignment over unimodal models

Table 8: Overview of multimodal model evaluation settings in brain encoding studies.

Study	Model Type	Stimulus Modality	Brain Data	Dataset	Instruction-Tuned
Doerig et al. (2022)	Vision-Language (CLIP)	Unimodal (Images)	fMRI	NSD	✗
Wang et al. (2023)	Vision-Language (CLIP)	Unimodal (Images)	fMRI	NSD	✗
Oota et al. (2022b)	Vision-Language (CLIP, VisualBERT, LXMERT)	Unimodal (Images)	fMRI	BOLD5000	✗
Popham et al. (2021)	Vision-Only CNNs vs. Vision-Language	Unimodal (Silent Videos)	fMRI	Gallant lab short video clips	✗
Tang et al. (2022)	non-instruction-tuned multimodal model (BridgeTower)	Unimodal (Silent Videos), Unimodal (listening stories)	fMRI	Gallant lab short video clips	✗
Oota et al. (2025a)	Instruction-tuned Image+Text MLLMs	Unimodal (Images)	fMRI	NSD	✓
Sartzetaki et al. (2025)	Image Recognition models, Action recognition models	Unimodal (Visual)	fMRI	Bold Moments Dataset	✗
Nakagi et al. (2024)	Language models (BERT, GPT-2, Lllama2, OPT)	Multimodal (Videos with audio)	fMRI	8.3 hours of video dataset	✗
Subramaniam et al. (2024)	non-instruction-tuned multimodal models (SLIP-CLIP, SimCLR, BLIP, BEIT)	Image frame-text pairs (Movies)	SEEG	AMMT	✗
Dong & Toneva (2023a)	non-instruction-tuned multimodal models (Merlore-serve)	Multimodal (Movies: Videos with audio)	fMRI	Neuromod Friends dataset	✗
Oota et al. (2025b)	non-instruction-tuned multimodal models (TVLT and ImageBind)	Multimodal (Movies: Videos with audio)	fMRI	Neuromod Movie10	✗
Our study	instruction-tuned video and audio MLLMs, in-context learning video and audio MLLMs	Multimodal (Movies: Videos with audio)	fMRI	Neuromod Movie10	✓

B Related work

Brain encoding using multimodal models. Our work is closely related to that of Conwell et al. (2024); Wang et al. (2023); Doerig et al. (2022); Tang et al. (2024); Nakagi et al. (2024); Dong & Toneva (2023b); Oota et al. (2025b), who proposed using multimodal model representations to study the contribution of brain alignment in unimodal and multimodal stimuli. The majority of brain encoding studies in using multimodal models focused on a single modality of input – vision alone (Conwell et al., 2024; Wang et al., 2023; Doerig et al., 2022; Wang et al., 2023; Tang et al., 2024; Nakagi et al., 2024; Gao et al., 2025). Recently, Dong & Toneva (2023b); Oota et al. (2022b) interpreted the effectiveness of multimodal Transformer language models in multimodal naturalistic stimuli. However, these studies focus on pretrained multimodal models which are not generic to tasks and lack the investigation of recent instruction-tuned models.

Task-based brain alignment. Task-specific fine-tuned Transformers align better with brain activity than pretrained models for text (Oota et al., 2022a; Aw & Toneva, 2023; Oota et al., 2024b), speech (Oota et al., 2023; Tuckute et al., 2023; Oota et al., 2024a), and vision (Wang et al., 2019;

Conwell et al., 2024), but require separate data and training per task, limiting generalization. Instruction tuning (Xu et al., 2023; Dai et al., 2023; Liu et al., 2024) is a scalable alternative, finetuning a single LLM across many tasks and surpassing task-specific models (Taori et al., 2023; Touvron et al., 2023; Dubey et al., 2024) while showing stronger brain alignment (Sun et al., 2023; Sun & Moens, 2023; Loong Aw et al., 2024) (App. B). Building on this, recent work aligns IT-MLLMs with brain data for text (Benara et al., 2024) and images (Oota et al., 2025a), again only under unimodal stimuli.

Our work is also closely related to that of Wang et al. (2019); Oota et al. (2022a); Aw & Toneva (2023); Sun et al. (2023) and Loong Aw et al. (2024), who propose using task-specific model representations to study the contribution of individual tasks to brain alignment. Wang et al. (2019) investigated 21 computer vision tasks to explore which vision tasks are more aligned with the brain while subjects engaged in viewing passive images. Similarly, Oota et al. (2022a) and Sun et al. (2023) explored 10 GLUE NLP tasks to study which NLP tasks are more brain-aligned during reading and listening to stories. More recent work by Loong Aw et al. (2024) uses instruction-tuned LLMs to investigate the effect of natural language instruction model representations on brain alignment across layers for language comprehension. Further, Oota et al. (2025a) use IT-MLLMs (image+text), using natural language instructions across diverse vision tasks to analyze their alignment with brain activity across layers during visual processing. However, these studies primarily focused on unimodal stimuli and thus do not fully capture the capabilities of multimodal instruction-tuned models under multimodal conditions. We complement these works by examining the impact of a wide range of IT-MLLMs—spanning video and audio-based models with text-based prompts—on their alignment with brain activity from multimodal stimuli.

C Detailed sub-ROIs of language, visual and auditory regions

The data covers seven brain regions of interest (ROIs) in the human brain with the following subdivisions: (i) early visual (EV: V1, V2, V3, V3B, and V4); (ii) object-related areas (LO1 and LO2); (iii) face-related areas (OFA), (iv) scene-related areas (PPA), (v) middle temporal (MT: MT, MST, LO3, FST and V3CD), (vi) late language regions, encompassing broader language regions: angular gyrus (AG: PFm, PGs, PGi, TPOJ2, TPOJ3), lateral temporal cortex (LTC: STSda, STSva, STGa, TE1a, TE2a, TGv, TGd, A5, STSdp, STSvp, PSL, STV, TPOJ1), inferior frontal gyrus (IFG: 44, 45, IFJa, IFSp) and middle frontal gyrus (MFG: 55b) (Baker et al., 2018; Milton et al., 2021; Desai et al., 2023). Fig. 7 shows flattened cortical surfaces for language-, visual- and auditory-selective regions displayed on the 'fsaverage' surface, used as the mask for all participants.

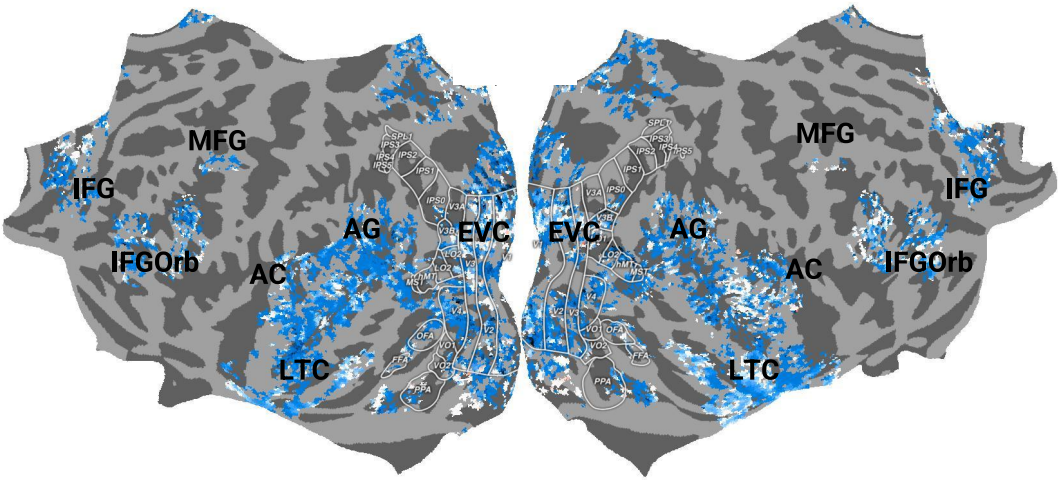


Figure 7: Flattened cortical surfaces for language-, visual- and auditory-selective regions displayed on the 'fsaverage' surface, used as the mask for all participants.

D CROSS-SUBJECT PREDICTION ACCURACY

We follow the method introduced by Schrimpf et al. (2021) to estimate how well brain activity in one individual can be predicted from others, using the Movie10 fMRI dataset. Starting with data from n participants (e.g., $n = 6$), for each subject $s \in$ ([1,6]) is chosen as the prediction target and the other five are used to predict this target, we use a voxel-wise encoding model (see Sec. 3) to predict one participant's response from others. For every combination, one participant was randomly chosen as the target, and the model was trained to predict their brain responses using data from the remaining $n-1$ participants. This gave us an average prediction score (correlation) for each voxel at each participant. To extrapolate to infinitely many humans and thus to obtain the highest possible (most conservative) estimate, as suggested by Schrimpf et al. (2021), we fit the equation $v = v_0 \times \left(1 - e^{-\frac{x}{\tau_0}}\right)$ where x is each subsample's number of participants, v is each subsample's correlation score and v_0 and τ_0 are the fitted parameters. This fitting was performed for each voxel independently with 100 bootstraps each to estimate the variance where each bootstrap draws x and v with replacement. The final ceiling value was the median of the per-voxel ceilings v_0.

We estimate cross-subject prediction accuracy in three settings: (i) training with *The Bourne supremacy* and testing with *Life* data, (ii) training with *The wolf of wall street* and testing with *Life* data, and (iii) training with both *The Bourne supremacy* and *The wolf of wall street* and testing with Life data. We present the average cross-subject prediction accuracy across voxels for the *Movie10 fMRI* dataset and across the three settings in Fig. 8.

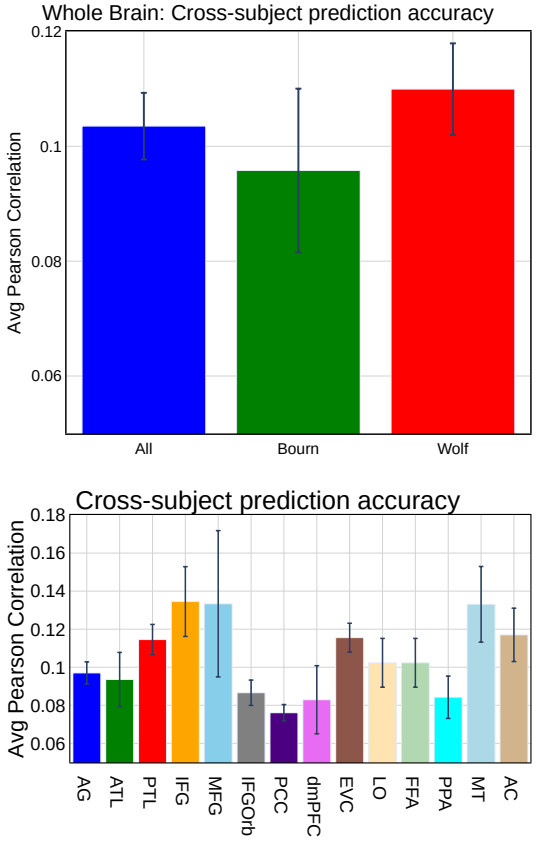


Figure 8: Cross-subject prediction accuracy: (top) across whole brain, (bottom) across language, visual and auditory regions.

E Implementation details for reproducibility

All feature extraction experiments were conducted on a machine equipped with an NVIDIA A100 GPU with 80 GB of GPU RAM, partitioned into two devices of 40 GB each. The voxelwise encoding models were trained on NVIDIA GeForce RTX 3050 GPU with 4GB of GPU RAM. We used banded ridge-regression with the following parameters: MSE loss function; L2-decay (λ) varied from 10^{-1} to 10^{3}; the best λ was chosen by tuning on validation data that comprised a randomly chosen 10% subset from the train set used only for hyper-parameter tuning.

F Statistical Significance

To determine if normalized predictivity scores are significantly higher than chance, we run a permutation test using blocks of 10 contiguous fMRI TRs (considering the slowness of hemodynamic response) rather than individual TRs. By permuting predictions 5000 times, we create an empirical distribution for chance performance, from which we estimate p-value of the actual performance. The choice of these specific permutation test configurations is based on established methodologies in previous research (Deniz et al., 2019; Reddy & Wehbe, 2021; Oota et al., 2024a).

For the category-level comparisons in Fig. 3, we compute, for each participant, normalized brain alignment at each model's best layer, averaged across instructions and across the models within each category. Categories are then compared with paired t-tests across the six participants ($n = 6$), so that each participant is an independent observation. Error bars show the standard error of the mean across participants. For task-conditioned comparisons within a model, alignment is averaged across participants separately for each task, and paired Wilcoxon signed-rank tests (Conover, 1999) are performed across the 13 matched task instructions ($n = 13$). The Benjamini–Hochberg false discovery rate (FDR) correction (Benjamini & Hochberg, 1995) is applied to all tests, which is appropriate because fMRI data are considered to have positive dependence (Genovese, 2000).

G Correlation Between Instruction Semantics and Model Representations

We conducted a comprehensive study of the robustness of semantic similarity.

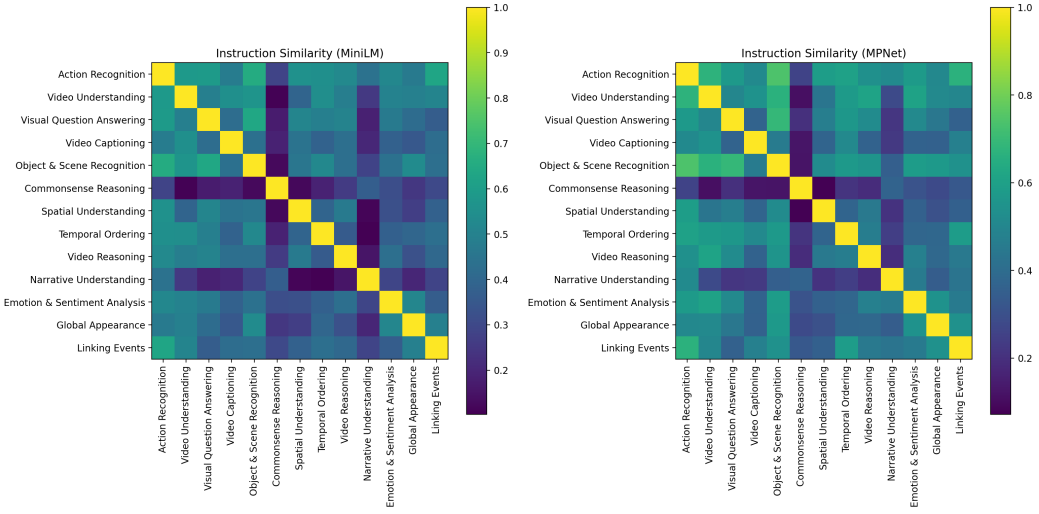


Figure 9: Instruction semantic similarity using text-embeddings.

Semantic similarity measurement. To validate that our task-specific instruction set contains semantically similar pairs, we perform following: (i) Using 13 video task-specific instructions, we first computed pairwise semantic similarity using two independent text embedding models (all-MiniLM-L6-v2 (Reimers & Gurevych, 2019) and MPNet (Song et al., 2020)). Both models produced highly consistent semantic similarity matrices (Pearson r = 0.94 between MiniLM and MPNet embed-

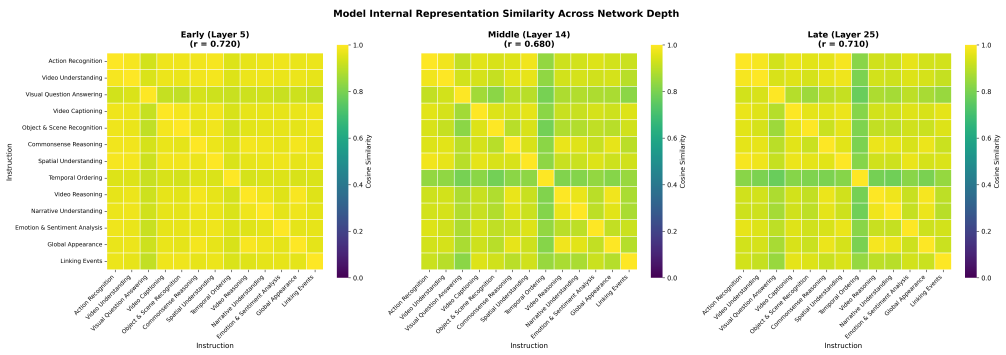


Figure 10: Qwen-2.5-VL-7B-Instruct: Model internal representational similarity across layer depth.

dings). (ii) Captures fine-grained semantic relationships between instruction texts. (iii) Example pairs: "Describe the video" vs. "Caption the video" (high semantic similarity). Fig. 9 shows pairwise semantic similarity between 13 instruction prompts computed using MiniLM embeddings (left) and MPNet (right). From Fig. 9 (left), we observe that the semantic similarity ranges from 0.15 to 0.85 (mean: 0.42 ± 0.18), with multiple pairs of high-similarity identified ($>$0.60):(Action Recognition vs. Video Understanding (0.68), Video Understanding vs. Visual Question Answering (0.65), Object & Scene Recognition vs. Action Recognition (0.68)). We also observe low semantic similarity pairs: (Commonsense Reasoning vs. Most others, Commonsense Reasoning vs. other tasks, Spatial Understanding vs. Emotional/Narrative). This confirms our instruction set contains the "semantically similar or equivalent instructions" for testing fine-grained task distinctions.

Internal representations of the IT MLLM. To measure how the model internally processes different task-specific instructions, we extracted hidden language states from the Qwen-2.5-VL-7B-Instruct model across all processing layers. We perform the following: (i) Extracted language hidden states across all 29 layers for each task-specific instruction. (ii) Used the same video input with varying instructions to isolate instruction effects. (iii) Analyzed three key layers: Early (Layer 5), Middle (Layer 14), and Late (Layer 25). (iv) While we present detailed results for three representative layers, we computed correlations and clustering metrics across all 29 layers to ensure findings are not artifacts of specific layer selection.

Fig. 10 shows pairwise similarity between 13 instruction prompts computed using three key layers. From Fig. 10, we observe that layer-specific differentiation patterns emerge: (i) In early layers, relatively uniform representation similarity suggests that initial encoding stage has limited task specialization, (ii) In middle layers, increased variation in similarity patterns indicates evidence of task-specific transformations, and (iii) In later layers, pattern similar to early layer but with subtle differences implies refinement of task-specific representations. Overall, the quantitative analysis from Table 1 reveal that middle layer shows greatest differentiation, while early and late layers show convergent patterns. While, we observe clear structure in model representations, a critical question remains: Do these patterns reflect semantic similarity or functional task requirements? To answer this, we next compare model representations with semantic similarity computed from text embeddings.

Correlation analysis: semantic similarity vs. model representations. For each layer, we computed correlations between: Semantic similarity (from text embeddings) and Model representation similarity (from hidden states) as follows: For each layer L in [1, 2, $\cdots$, 29]: (i) Extract upper triangle of semantic similarity matrix (78 pairs), (ii) Extract upper triangle of model similarity matrix (78 pairs), (iii) Compute Pearson correlation between the two sets, (iv) Compute Spearman correlation for robustness, and (v) Test significance against random baseline

We compute both (a) the semantic similarity between instructions using text embedding models, and (b) the similarity of internal model representations (e.g., cosine similarity between layer vectors). A strong positive correlation between these measures would demonstrate that instruction tuning enhances the model's sensitivity to fine-grained semantic distinctions. Contrary to the hypothesis that surface semantic similarity should drive internal similarity, as shown in Table 9, we observe near-zero to modestly positive coupling (Pearson r $\approx$ 0.04–0.18; Spearman $\rho \approx$ 0.15–0.27), with several mid/late layers show uncorrected Spearman p$<$0.05, but no layer survives FDR correction across

Table 9: Correlation across network depth. Coupling between instruction semantic similarity and instruction-conditioned representation similarity.

Layer range	Pearson r (mean $\pm$ sd)	p-value range	Spearman ρ (mean $\pm$ sd)	Significance
Early (L1–L10)	0.066 ± 0.015	0.497–0.717	0.184 ± 0.022	No individual layer sig. after FDR
Middle (L11–L20)	0.126 ± 0.009	0.219–0.330	0.199 ± 0.011	Some uncorrected $p < 0.05$; none FDR-sig
Late (L21–L28)	0.149 ± 0.018	0.109–0.240	0.247 ± 0.013	Several uncorrected $p < 0.05$; none FDR-sig
Overall (L1–L29)	0.113 ± 0.041	0.109–0.717	0.207 ± 0.034	No FDR-significant effects

29 layers (Benjamini–Hochberg). This pattern is highly consistent throughout the network depth, indicating that instruction tuning does not enhance semantic sensitivity; instead, IT representations show task-conditioned clustering

Statistical validation (Mantel permutation). We assess whether instruction semantics predict representation similarity using a Mantel permutation test (10,000 label shuffles) per layer, with FDR correction across layers. Pearson r and Spearman ρ are the raw matrix correlations; Mantel z and q report significance. From Table 10, we find that Mantel tests show no significant association between semantic and representation spaces across layers, indicating representations are not organized by surface semantic similarity.

Table 10: **Mantel permutation test (semantic vs. representation similarity) for IT models of the three matched pairs.** 10,000 label permutations; FDR applied across layers.

Model	Layer	Pearson r	Spearman ρ	Mantel z	q (FDR)
Qwen-2.5-VL-7B-Instruct	L5	0.077	0.210	0.65	0.482
	L14	0.130	0.201	1.12	0.376
	L25	0.143	0.237	1.23	0.338
Qwen-2-VL-7B-Instruct	L5	0.091	0.184	0.78	0.451
	L14	0.116	0.213	1.01	0.392
	L25	0.158	0.246	1.36	0.318
InternVL3-8B-Instruct	L5	0.052	0.162	0.44	0.563
	L14	0.101	0.196	0.87	0.427
	L25	0.126	0.221	1.08	0.361

Clustering Quality Analysis. We perform clustering analysis to determine whether the model organizes instructions by semantic similarity or functional task categories.

Table 11: **Task-based vs. semantic clustering alignment (ARI) for three matched pairs.** ARI(Sem) - agreement between representation clusters and instruction-semantic clusters; Δ = ARI(Task) $-$ ARI(Sem). Reported descriptively; task-conditioned structure is assessed with the within-video permutation test.

		ARI (Sem)			Δ ARI		
Pair	Model	L5	L14	L25	L5	L14	L25
Qwen-2.5-VL-7B	IT	-0.132	-0.132	-0.132	$+0.129$	$+0.142$	$+0.174$
	ICL	$+0.421$	$+0.327$	$+0.367$	-0.442	-0.291	-0.349
Qwen-2-VL-7B	IT	-0.114	-0.125	-0.114	$+0.083$	$+0.151$	$+0.132$
	ICL	$+0.342$	$+0.364$	$+0.315$	-0.318	-0.376	-0.284
InternVL3-8B	IT	-0.114	-0.114	-0.128	$+0.096$	$+0.118$	$+0.165$
	ICL	$+0.312$	$+0.343$	$+0.383$	-0.271	-0.337	-0.402

Note: ARI > 0 indicates clustering aligns with task-based labels, and < 0 for semantic further indicates that functional labels align better with representation geometry.

From Table 11, we make the following key findings: (i) Task-based categories consistently outperform semantic clusters across all metrics and layers. (ii) ARI advantage: Δ = +0.129 to +0.174 (strong functional alignment). (iii) Middle layer specialization: mid/later layers shows strongest functional differentiation (ARI Δ = +0.423)

Task-conditioned organization across layers. ARI measures agreement between representation clusters and a given grouping (task categories or instruction-semantic clusters). Across the three pairs, ICL representation clusters align with instruction-semantic groupings (ARI(Sem) $= 0.31$–0.42), whereas IT clusters do not (ARI(Sem) ≈ -0.12); ARI with task categories is near zero for both. We therefore use ARI only as descriptive evidence that IT representations depart from semantic organization. The primary statistical evidence for task-conditioned structure is the within-video permutation analysis (Table 6 in the main text): Δ_{task} is significant at all layers for all three IT models and at no layer for the ICL models, and it increases with depth in all three IT models (Spearman $\rho = 0.84$–0.91, all $p < 0.001$).

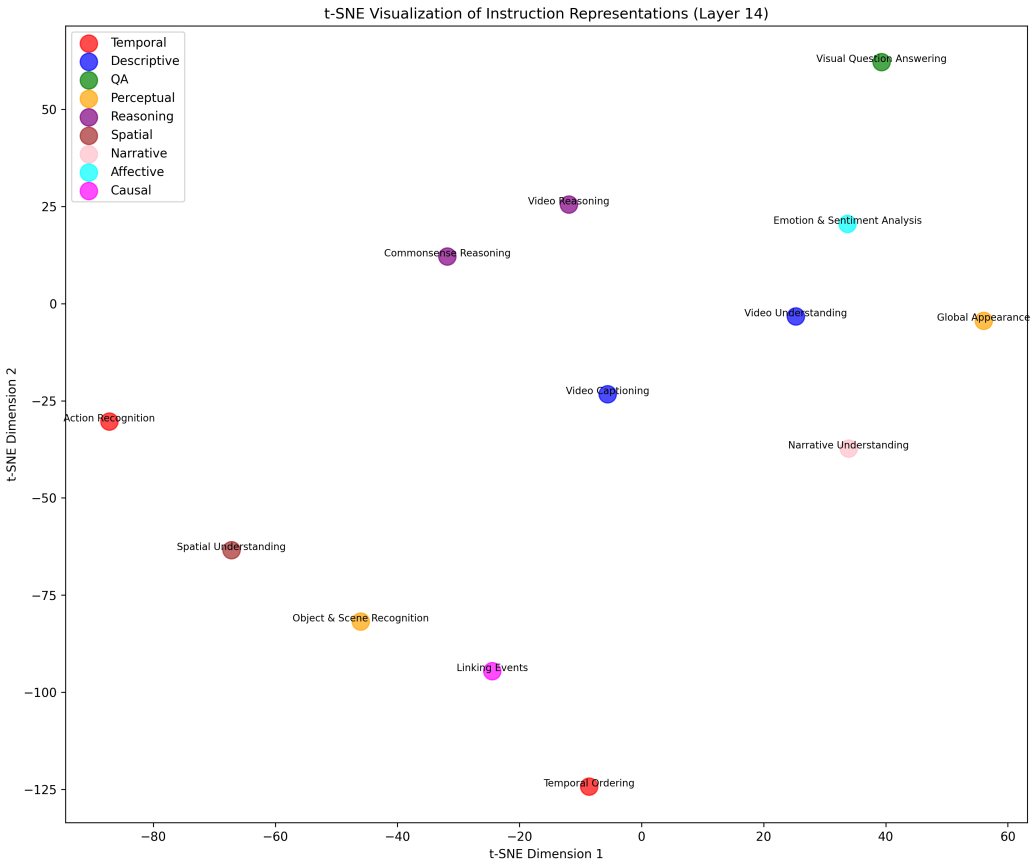


Figure 11: Clustering instruction task-specific representations.

Visualization: t-SNE Projections (illustrative only) We performed t-SNE dimensionality reduction to visualize how the model organizes instruction task-specific representations in each layer (see Fig. 11, attached). We emphasize that t-SNE is used for visualization only; conclusions rely on ARI/silhouette and Mantel statistics. From Fig. 11, we make the following observations: (1) Instructions cluster by task-based labels rather than semantic similarity. (2) Clear spatial separation between: (i) Perceptual prompts(object recognition, appearance) to Right region, (ii) Reasoning prompts(commonsense, video reasoning) to Center region, (iii) Temporal-causal prompts(action recognition, event linking) to Bottom region, (iv) Descriptive prompts(video understanding, captioning, narrative) to Distributed pattern. (3) Semantically similar instructions (e.g., "Describe video" vs. "Caption video") are spatially separated. (4) Functional categories show consistent grouping patterns across layers.

Table 12: IT vs. ICL models: Representational similarity and clustering metrics. *** p<0.001, FDR-significant across 29 layers; ns = not significant after FDR.

Metric	Model	Layer 5	Layer 14	Layer 25	Winner	Magnitude
Semantic correlation	IT (Qwen-2.5-VL-7B-Instruct)	Pearson $r = 0.077$ (ns)	0.130 (ns)	0.143 (ns)	—	Weak
	ICL (Qwen-2.5-VL-7B)	**0.690*** **	**0.680*** **	**0.777*** **	**ICL**	$\sim$4.2$\times$ stronger
Adjusted Rand Index (label alignment)	IT Δ (Func $-$ Sem)	**+0.129**	**+0.142**	**+0.174**	**Functional**	Weak
	ICL Δ (Func $-$ Sem)	-0.442	-0.291	-0.349	**Semantic**	Very Strong

H Representational Differences Between IT and ICL Models

ICL models can follow zero-shot prompting i.e. in-context prompt, whereas Instruction-tuning adds a supervised signal. Below we describe the representational differences observed between IT and ICL models.

To understand the difference in the working of IT and ICL models, we perform additional analysis. We compared instruction-tuned (IT) and in-context learning (ICL) models to identify differences in representational structure. For the 13 tasks, we first compute a 13×13 semantic-similarity matrix using MiniLM embeddings. We then perform Representational Similarity Analysis (RSA) by computing correlations between the upper triangles of the 13×13 semantic-similarity matrix and the corresponding representation-similarity matrix (same videos, same pipeline), per layer.

We find that for the instruction-tuned model (Qwen-2.5-VL-7B-Instruct), the correlation between instruction semantic similarity and internal representation similarity is weak across layers. (e.g., best layer L28: Pearson r=0.183, p=0.109; Spearman ρ=0.266, p=0.018); no layer remains significant after FDR correction across 29 layers. This indicates that IT representations are less coupled to surface prompt semantics than ICL representations, consistent with task-conditioned representational structure that diverges from instruction-text wording.

ICL (Qwen-2.5-VL-7B): RSA correlation r≈0.78 strong coupling to prompt semantics to shallow text-proximal matching. IT (Qwen-2.5-VL-7B-Instruct): RSA correlation r=0.183 (best layer, ns after FDR correction) weak coupling to prompt semantics to task-conditioned representational structure that diverges from instruction-text wording

- Adjusted Rand Index (ARI) measures agreement between predicted clusters and ground-truth labels by counting pairwise agreements, adjusted for chance, with values ranging from -1 (worse than random) to +1 (perfect agreement). Positive ARI indicates clustering aligns with task-based grouping, Negative ARI clustering aligns with semantic-based grouping (Hubert & Arabie, 1985).

Key findings: ICL vs IT. Across the three matched pairs, ICL representations are strongly coupled to instruction-text semantics (RSA $r = 0.69$–0.78 at L25; FDR $q < 0.001$), whereas IT representations are only weakly coupled ($r = 0.13$–0.16; no layer significant after FDR correction), a 4.4–5.6$\times$ difference (Table 5). Consistent with this, ICL representation clusters align with instruction-semantic groupings (ARI $= 0.31$–0.42), whereas IT clusters do not (ARI ≈ -0.12; Table 11); we treat ARI as descriptive only. The within-video permutation analysis provides the primary evidence for task-conditioned structure: Δ_{task} is significant at every layer for all three IT models and at no layer for the ICL models, and it increases with depth in all three IT models (Spearman $\rho = 0.84$–0.91, all $p < 0.001$; Table 6).

ICL representations are more strongly coupled to instruction-text semantics, whereas IT representations show weaker semantic coupling and significant task-conditioned structure throughout the network, without a significant monotonic increase with depth. This is consistent with prior findings that instruction tuning leads to task-specific representations supporting downstream task performance (Zhu et al., 2025), and we extend this by showing that these representations also yield improved brain alignment.

I Effectiveness of instruction-tuned video MLLMs vs audio MLLMs vs multimodal vs unimodal representations for various brain regions

Fig. 13 show average normalized brain alignment of instruction-tuned video MLLMs vs instruction-tuned audio MLLMs vs multimodal and unimodal models across several ROIs (AG, ATL, PTL, IFG, MFG, IFGOrb, PCC and dmPFC) of language region. Fig. 14 show the same for visual, auditory and motor regions.

Fig. 15 shows task-wise best-voxel fractions in key higher-order language and visual ROIs to assess how instructions differentially engage them. As shown in Fig. 15, restricting the analysis to spatial-visual tasks (Spatial Understanding, Temporal Ordering, Object & Scene Recognition) reveals clean dissociations across early visual ROIs. PPA is dominated by Object & Scene Recognition (48.5%) and Spatial Understanding (41.0%), consistent with its established role in scene and place processing. OFA shows a balanced profile led by Object & Scene Recognition (38.1%), consistent with its role in low-level object recognition including faces. LOC is dominated by Spatial Understanding (51.2%), a richer task than bare recognition, consistent with LOC's role in object and scene process-

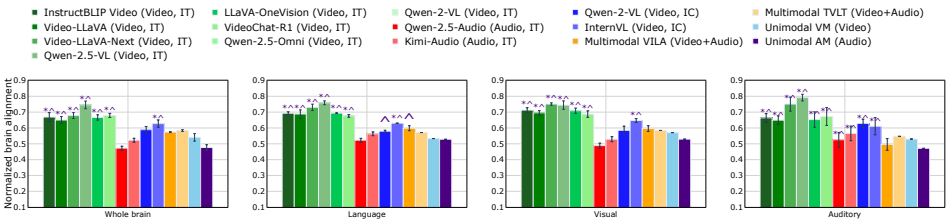


Figure 12: Average normalized brain alignment of instruction-tuned video MLLMs vs instruction-tuned audio MLLMs vs in-context learning video MLLMs vs multimodal and unimodal models across whole brain, language, visual and auditory regions. Error bars indicate the standard error of the mean across participants. $*$ implies that instruction-tuned MLLM embeddings are significantly better than multimodal models and $\wedge$ means that instruction-tuned MLLM embeddings are significantly better than unimodal models with p≤ 0.05. IT: Instruction-tuned, IC: In-context learning

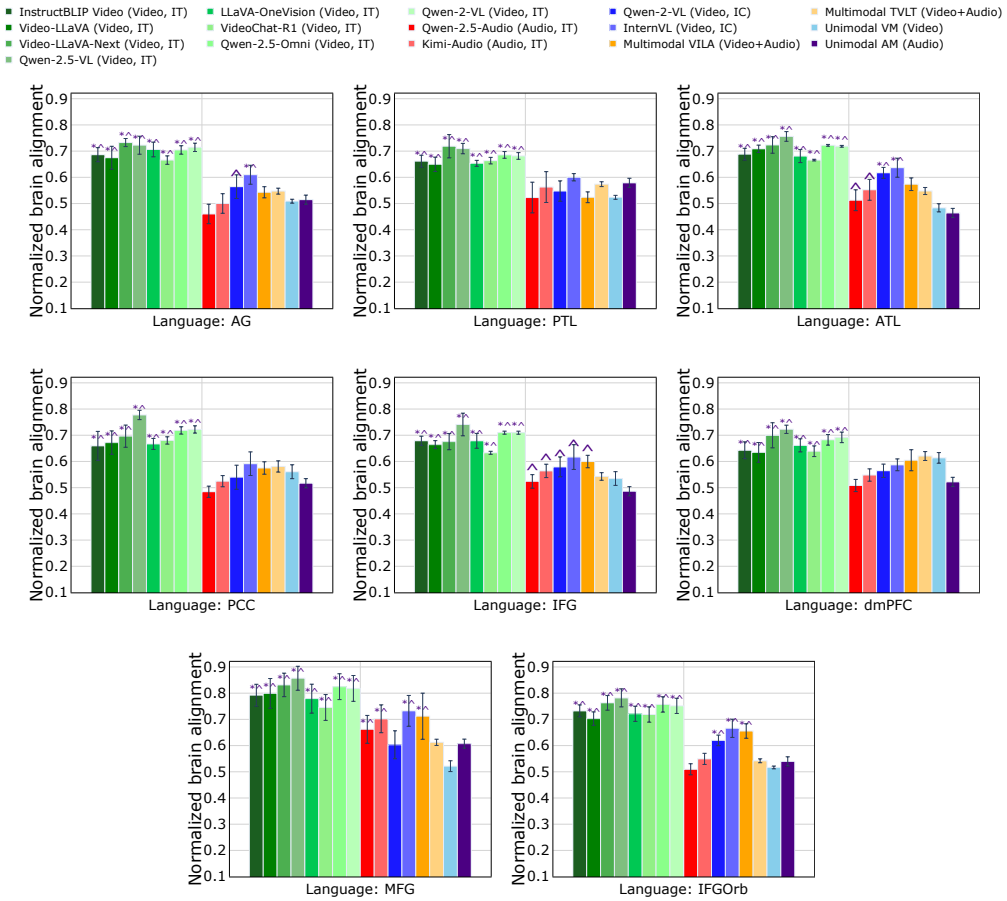


Figure 13: **Language ROIs**Average normalized brain alignment of instruction-tuned video MLLMs vs instruction-tuned audio MLLMs vs multimodal and unimodal models across several ROIs (AG, ATL, PTL, IFG, MFG, IFGOrb, PCC and dmPFC) of language region. Error bars indicate the standard error of the mean across participants. $*$ implies that instruction-tuned MLLM embeddings are significantly better than multimodal models and $\wedge$ means that instruction-tuned MLLM embeddings are significantly better unimodal models with p≤ 0.05.

ing under spatial layout reasoning. These findings align with classical neuroanatomical findings on functional specialization across early visual cortex.

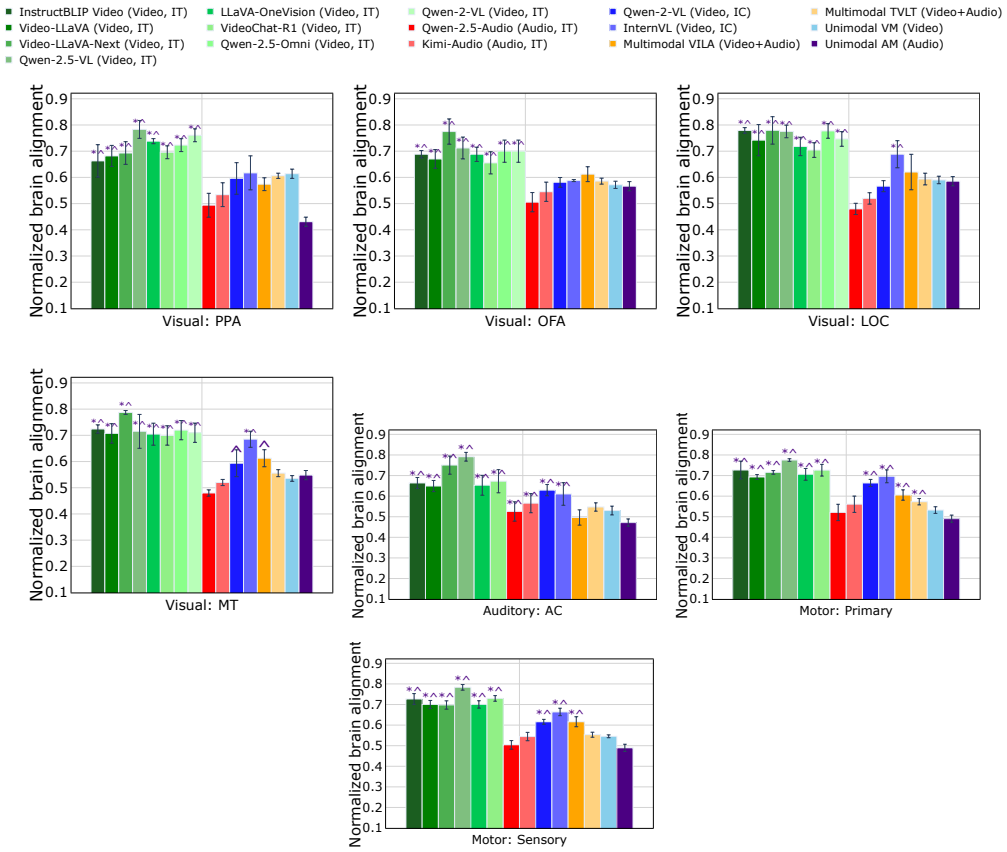


Figure 14: **Visual and Auditory ROIs** Average normalized brain alignment of instruction-tuned video MLLMs vs instruction-tuned audio MLLMs vs multimodal and unimodal models across several ROIs of visual cortex (PPA, OFA, LOC, MT), Auditory cortex (AC), and Motor Area (PMA and SMA). Error bars indicate the standard error of the mean across participants. $*$ implies that instruction-tuned MLLM embeddings are significantly better than multimodal models and $\wedge$ means that instruction-tuned MLLM embeddings are significantly better unimodal models with $p \leq 0.05$.

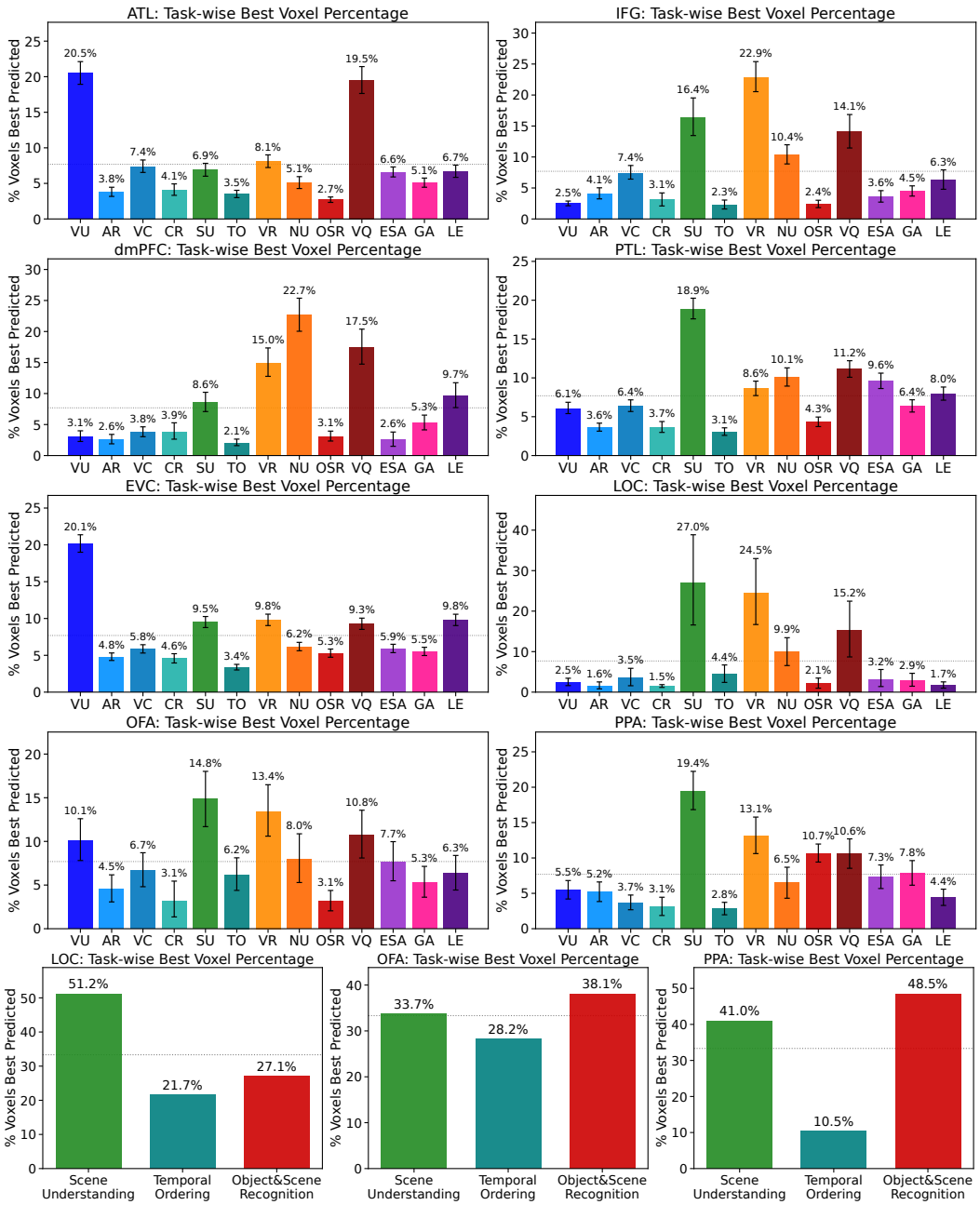


Figure 15: Percentage of best predicted voxels per task for an instruction-tuned video MLLM across three language ROIs: ATL, IFG and dmPFC.

Table 13: **ROI × task-preference distributions.** Percentage of voxels best predicted by each of the 13 task instructions, with 95% confidence intervals estimated using 10,000 within-ROI voxel-bootstrap resamples. Results use Qwen-2.5-VL-7B-Instruct, Layer 28, across six participants. The equal-preference reference is 7.69% (100/13). Bold values have confidence intervals whose lower bound exceeds this reference. Because voxels are spatially correlated, these intervals quantify within-ROI uncertainty and are interpreted alongside participant-level reproducibility.

Task	AG	ATL	PTL	IFG	MFG	IFGOrb
VU	3.61 [2.96, 4.30]	**20.50 [18.93, 22.12]**	6.11 [5.40, 6.86]	2.50 [2.13, 2.89]	2.40 [1.01, 3.90]	6.03 [4.84, 7.29]
AR	1.65 [1.35, 1.97]	3.80 [3.17, 4.46]	3.64 [3.14, 4.17]	4.10 [3.24, 5.03]	1.40 [0.14, 2.83]	3.09 [2.45, 3.83]
VC	4.39 [3.59, 5.23]	7.40 [6.54, 8.29]	6.42 [5.67, 7.17]	7.40 [6.43, 8.63]	6.30 [4.09, 8.68]	**14.61 [12.49, 16.77]**
CR	4.07 [3.33, 4.85]	4.10 [3.32, 4.93]	3.67 [2.98, 4.38]	3.10 [2.10, 4.20]	2.40 [0.49, 4.89]	2.34 [1.30, 3.51]
SU	**14.75 [13.52, 16.02]**	6.90 [6.01, 7.82]	**18.91 [17.60, 20.26]**	**16.40 [13.46, 19.53]**	1.50 [0.52, 2.63]	**10.30 [8.39, 12.35]**
TO	2.10 [1.69, 2.54]	3.50 [3.00, 4.03]	3.07 [2.60, 3.57]	2.30 [1.61, 3.05]	2.60 [1.48, 3.97]	4.47 [3.48, 5.57]
VR	**9.50 [8.53, 10.51]**	8.10 [7.23, 9.01]	**8.63 [7.72, 9.58]**	**22.90 [20.54, 25.39]**	9.80 [6.00, 14.02]	8.43 [6.82, 10.08]
NU	**20.27 [18.81, 21.77]**	5.10 [4.28, 5.95]	**10.10 [8.94, 11.29]**	**10.40 [8.88, 11.97]**	**12.20 [7.95, 16.80]**	**18.57 [16.14, 21.12]**
OSR	2.75 [2.27, 3.26]	2.70 [2.33, 3.09]	4.33 [3.73, 4.96]	2.40 [1.83, 3.02]	3.90 [2.31, 5.92]	7.52 [6.06, 9.03]
VQ	**14.80 [13.57, 16.08]**	**19.50 [17.64, 21.42]**	**11.15 [10.08, 12.21]**	**14.10 [11.45, 16.85]**	**20.10 [16.75, 23.96]**	8.57 [7.01, 10.20]
ESA	6.87 [6.06, 7.72]	6.60 [5.91, 7.31]	**9.60 [8.63, 10.61]**	3.60 [2.72, 4.57]	**26.40 [21.04, 32.08]**	5.08 [3.97, 6.30]
GA	4.19 [3.57, 4.84]	5.10 [4.46, 5.75]	6.39 [5.60, 7.18]	4.50 [3.69, 5.34]	3.20 [1.43, 5.49]	7.07 [5.70, 8.51]
LE	**11.09 [10.04, 12.19]**	6.70 [5.85, 7.57]	7.97 [7.13, 8.82]	6.30 [4.79, 7.92]	7.80 [4.89, 11.09]	3.90 [3.04, 4.87]

Task	PCC	dmPFC	EVC	LOC	OFA	PPA
VU	5.81 [4.84, 6.85]	3.10 [2.27, 3.99]	**20.10 [19.00, 21.37]**	2.50 [1.61, 3.43]	**10.11 [7.81, 12.60]**	5.45 [4.20, 6.81]
AR	2.60 [2.10, 3.13]	2.60 [1.89, 3.41]	4.80 [4.30, 5.32]	1.60 [0.81, 2.52]	4.52 [3.07, 6.17]	5.18 [3.84, 6.61]
VC	4.70 [3.86, 5.57]	3.80 [3.04, 4.63]	5.85 [5.31, 6.44]	3.50 [1.58, 5.88]	6.66 [4.81, 8.70]	3.67 [2.68, 4.76]
CR	5.10 [4.06, 6.25]	3.90 [2.64, 5.28]	4.57 [3.96, 5.22]	1.50 [1.16, 1.85]	3.12 [1.35, 5.45]	3.11 [1.87, 4.45]
SU	**11.30 [9.69, 12.94]**	8.60 [7.09, 10.20]	**9.53 [8.77, 10.27]**	**27.00 [16.58, 38.85]**	**14.83 [11.69, 18.02]**	**19.44 [16.83, 22.24]**
TO	3.30 [2.63, 4.04]	2.10 [1.60, 2.65]	3.38 [3.01, 3.77]	4.40 [2.40, 6.71]	6.19 [4.39, 8.12]	2.80 [1.97, 3.73]
VR	6.60 [5.67, 7.57]	**15.00 [12.77, 17.35]**	**9.81 [9.04, 10.59]**	**24.50 [16.69, 32.98]**	**13.44 [10.60, 16.49]**	**13.11 [10.62, 15.78]**
NU	**28.30 [26.20, 30.44]**	**22.70 [20.04, 25.37]**	6.18 [5.62, 6.76]	9.90 [6.56, 13.44]	7.98 [5.29, 10.87]	6.48 [4.31, 8.69]
OSR	4.60 [3.84, 5.42]	3.10 [2.34, 3.94]	5.28 [4.74, 5.85]	2.10 [0.96, 3.47]	3.11 [2.05, 4.38]	**10.67 [9.42, 11.97]**
VQ	**9.20 [8.06, 10.38]**	**17.50 [14.74, 20.38]**	**9.27 [8.52, 10.05]**	**15.20 [8.69, 22.46]**	**10.76 [8.10, 13.57]**	**10.58 [8.54, 12.70]**
ESA	4.50 [3.74, 5.32]	2.60 [1.49, 3.79]	5.92 [5.36, 6.50]	3.20 [1.38, 5.59]	7.66 [5.49, 9.97]	7.30 [5.68, 9.01]
GA	6.40 [5.36, 7.50]	5.30 [4.14, 6.52]	5.51 [4.97, 6.09]	2.90 [1.46, 4.65]	5.29 [3.61, 7.14]	7.85 [6.15, 9.63]
LE	7.60 [6.52, 8.73]	**9.70 [7.73, 11.74]**	**9.80 [9.04, 10.58]**	1.70 [0.98, 2.53]	6.31 [4.43, 8.39]	4.38 [3.29, 5.59]

Task abbreviations: VU = Video Understanding; AR = Action Recognition; VC = Video Captioning; CR = Commonsense Reasoning; SU = Spatial Understanding; TO = Temporal Ordering; VR = Video Reasoning; NU = Narrative Understanding; OSR = Object and Scene Recognition; VQ = Visual Question Answering; ESA = Emotion and Sentiment Analysis; GA = Global Appearance; LE = Linking Events.

Table 14: **Participant-level reproducibility of ROI task preferences.** For each participant, the table reports the task instruction predicting the largest proportion of voxels within each ROI. Agreement indicates how many participants reproduce the group-level winner. The voxel-agreement column reports the percentage of voxels assigned the same task in at least four of six participants.

ROI	S1	S2	S3	S4	S5	S6	Group winner	Agreement	Voxels agreeing in ≥5/6	Unanimous
AG	NU	VQ	NU	NU	NU	NU	**NU**	5/6	32.8%	–
ATL	ESA	NU	VU	VU	VU	VU	**VU**	4/6	32.6%	–
PTL	SU	GA	SU	VQ	SU	SU	**SU**	4/6	32.2%	–
IFG	SU	VR	VR	VR	VR	VR	**VR**	5/6	34.5%	–
MFG	ESA	ESA	NU	ESA	ESA	ESA	**ESA**	5/6	26.5%	–
IFGOrb	NU	NU	VC	NU	NU	NU	**NU**	5/6	36.0%	–
PCC	NU	GA	NU	NU	NU	NU	**NU**	5/6	30.9%	–
dmPFC	NU	NU	NU	VQ	NU	NU	**NU**	5/6	28.1%	–
EVC	VU	VU	VU	VU	VU	VU	**VU**	6/6	29.0%	Yes
LOC	VU	VU	VU	VR	VU	VU	**VU**	5/6	21.8%	–
OFA	SU	SU	ESA	VQ	SU	SU	**SU**	4/6	30.3%	–
PPA	VR	SU	SU	SU	SU	SU	**SU**	5/6	28.6%	–

Summary: The group-level winner is reproduced by at least five of six participants in 9 of 12 ROIs, with a mean agreement of 4.83 participants. EVC shows unanimous agreement across all six participants.

Table 15: **Participant-level reproducibility of Narrative Understanding (NU) versus Video Understanding (VU) contrasts.** Each cell indicates whether the participant-level direction agrees with the group-level contrast.

Directional contrast	S1	S2	S3	S4	S5	S6	Consistent participants	Group direction
NU > VU in AG	Yes	No	Yes	Yes	Yes	Yes	**5/6**	Yes
NU > VU in IFGOrb	Yes	Yes	No	Yes	Yes	Yes	**5/6**	Yes
NU > VU in PCC	Yes	No	Yes	Yes	Yes	Yes	**5/6**	Yes
NU > VU in dmPFC	Yes	Yes	Yes	No	Yes	Yes	**5/6**	Yes
VU > NU in ATL	No	No	Yes	Yes	Yes	Yes	4/6	Yes
VU > NU in EVC	Yes	Yes	Yes	Yes	Yes	Yes	**6/6**	Yes

Summary: Five of the six directional contrasts are reproduced in at least five of six participants. The VU > NU contrast in EVC is reproduced by all six participants, whereas the corresponding ATL contrast is less consistent and is therefore not treated as a reproducible regional dissociation.

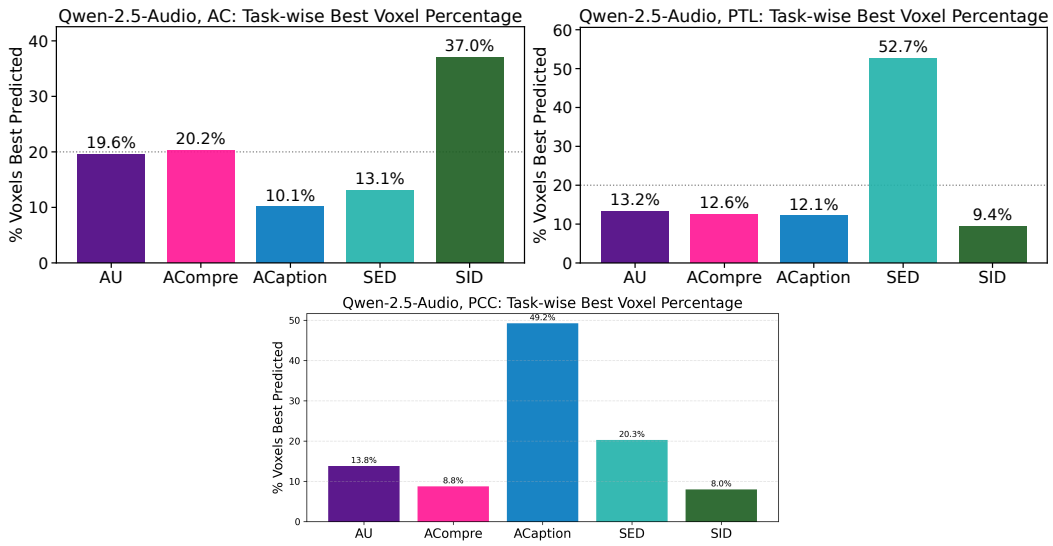
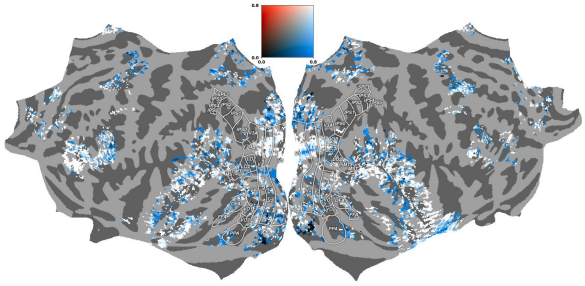


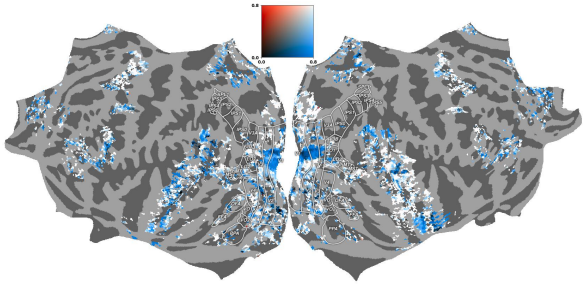
Figure 16: Percentage of best predicted voxels per task for an instruction-tuned audio MLLM across three language ROIs: AC, PTL and PCC.

J CONTRASTING INSTRUCTION-TUNED VIDEO MLLMS WITH IN-CONTEXT LEARNING VIDEO MLLMS

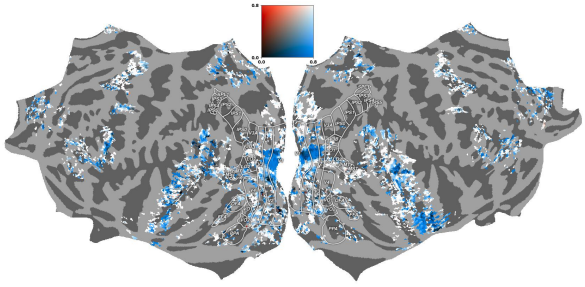
We present contrast of brainmaps to display the average normalized brain alignment across voxels for the instruction-tuned video MLLMs versus the in-context learning video MLLMs in Figures 17, and 18. The results show that instruction-tuned video MLLMs consistently achieve significantly higher alignment across all brain voxels.



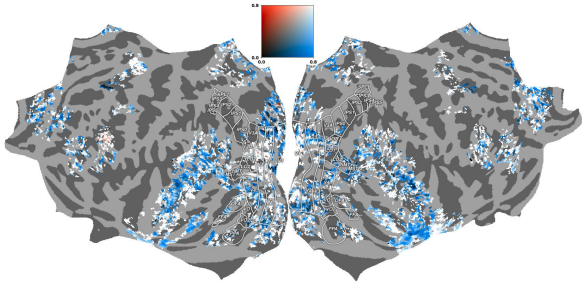
(a) Subject-01



(b) Subject-02

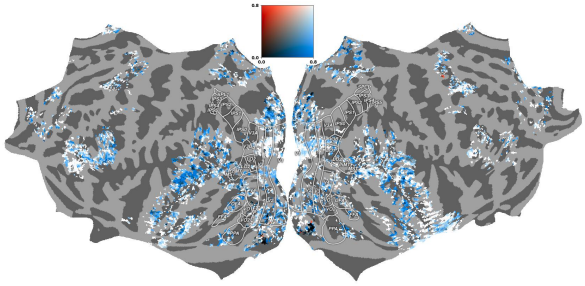


(c) Subject-03

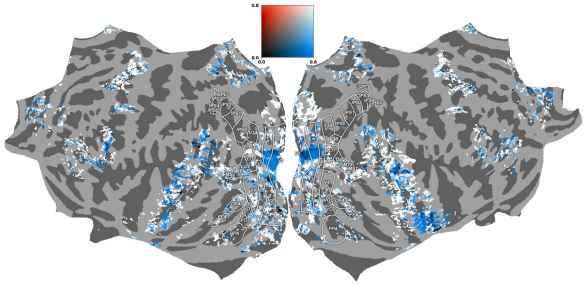


(d) Subject-05

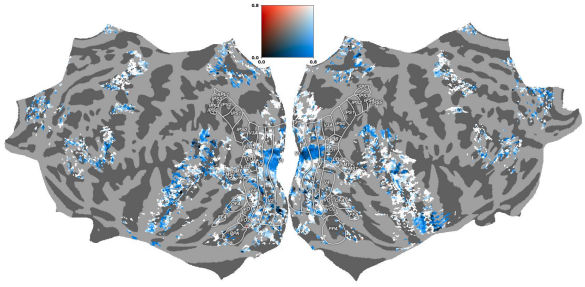
Figure 17: Qwen-2.5-VL-IT vs. InternVL: Contrast of estimated cross-subject prediction accuracy for all participants for the naturalistic movie watching. Pearson correlation scores for each voxel in each subject are projected onto the subject's flattened cortical surface. Blue and Red voxels depict higher prediction accuracy estimates during instruction-tuned video MLLM and in-context learning video MLLM (InternVL), respectively. Voxels that have similar cross-subject prediction accuracy appear white. Here, middle frontal gyrus (MFG), inferior frontal gyrus (IFG), inferior frontal gyrus orbital (IFGOrb), angular gyrus (AG), and lateral temporal cortex (LTC) are late language regions, EVC denotes early visual cortex and AC denotes auditory cortex.



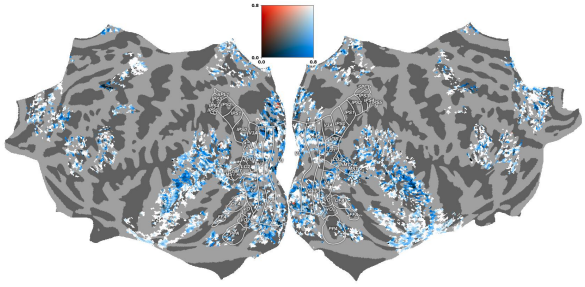
(a) Subject-01



(b) Subject-02



(c) Subject-03



(d) Subject-05

Figure 18: Qwen-2.5-VL-IT vs. Qwen-2.5-VL: Contrast of estimated cross-subject prediction accuracy for all participants for the naturalistic movie watching. Pearson correlation scores for each voxel in each subject are projected onto the subject's flattened cortical surface. Blue and Red voxels depict higher prediction accuracy estimates during instruction-tuned video MLLM and in-context learning video MLLM (Qwen-2.5-VL), respectively. Voxels that have similar cross-subject prediction accuracy appear white. Here, middle frontal gyrus (MFG), inferior frontal gyrus (IFG), inferior frontal gyrus orbital (IFGOrb), angular gyrus (AG), and lateral temporal cortex (LTC) are late language regions, EVC denotes early visual cortex and AC denotes auditory cortex.

K BRAIN MAPS FOR TASK-SPECIFIC INSTRUCTIONS

Figures 19 and 20 show brain maps for InstructBLIPVideo, Video-LLaVA, LLaVA-NeXT-Video, LLaVA-OneVision and VideoChat-R1 for video tasks for average normalized brain predictivity across subjects where the voxel color codes are projected onto the flattened cortical surface of the 'fsaverage' subject. The color-scheme corresponding to each instruction is also reported. We make the following observations: (i) Video understanding exhibits the strongest alignment across the whole brain. (ii) Tasks such as spatial understanding, narrative understanding, and visual question answering show higher alignment in language-related regions, including the angular gyrus, posterior temporal lobe, and visual regions. (iii) Higher-order language regions in the frontal cortex are predominantly identified by the video understanding task, with a smaller proportion of voxels also activated by video Reasoning and temporal ordering tasks.

Fig. 21 shows brainmap for audio instruction-tuned MLLM (Kimi-Audio) where the predictions are average across subjects. Here, the voxel color codes are projected onto the flattened cortical surface of the 'fsaverage' subject. The figure shows a clear distinction between different audio tasks.

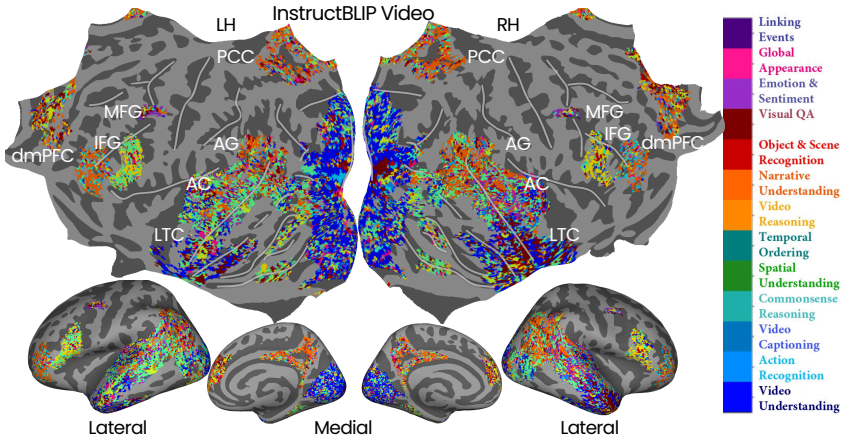
L BRAIN MAPS SHOWING LAYER-WISE DETAILS FOR VIDEO INSTRUCTION-BASED MLLMS

To examine whether IT-MLLMs reflect the brain's hierarchy of information processing across layers, we analyze the voxels as follows. For each voxel, we select the layer that results in the highest normalized brain alignment and apply a color code for the 29/33 layers across the various MLLMs. Fig. 22 presents brain maps for four video MLLMs, where the voxels with their corresponding color codes are projected onto the flattened cortical surface of the 'fsaverage' subject.

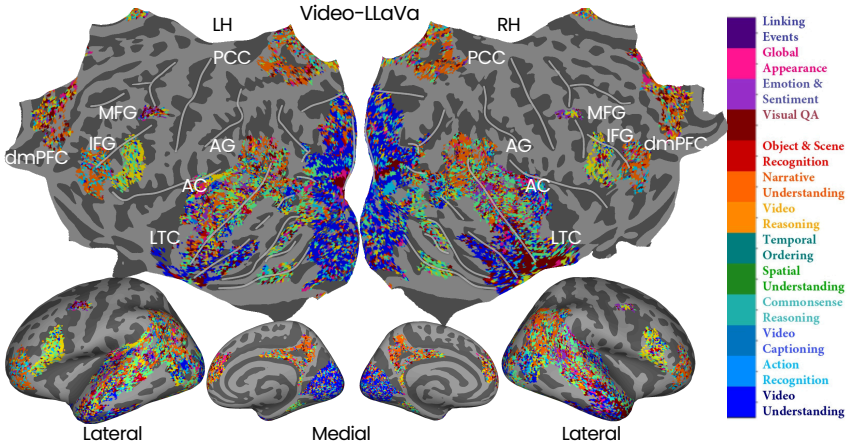
M VALIDATION OF "HIERARCHICAL CORRESPONDENCE" BETWEEN MODEL LAYERS AND BRAIN REGIONS

We compare hierarchical correspondence between model layers and brain regions in two settings: (i) Scaling the instruction-tuned model with correct video prompt: Qwen-2.5-vl-3B Instruct vs. Qwen-2.5-vl-7B Instruct, (ii) an instruction-tuned model with a non-natural language prompt, e.g.: ##### #### #### ##### this does not provide any meaning while passing video as input. Similar to Qwen-2.5-VL-7B Instruct model, we perform qualitative analysis by considering each voxel is color coded with the MLLM layer number (out of 29) that led to the highest normalized brain alignment. Fig 24 shows the resulting brain maps for the 3B-Instruct model (a) and for the 7B-Instruct model with the non-language prompt (b). We make the following observations: (i) Non-language control (random prompt): There is no hierarchy of information processing observed across layers, where the brain prediction is dominated by early layers across cortex (ii) Instruction condition (3B model): The hierarchical pattern still remains the same i.e., a systematic early to mid/late gradient across cortex, i.e., the same hierarchical pattern we observed with the larger model.

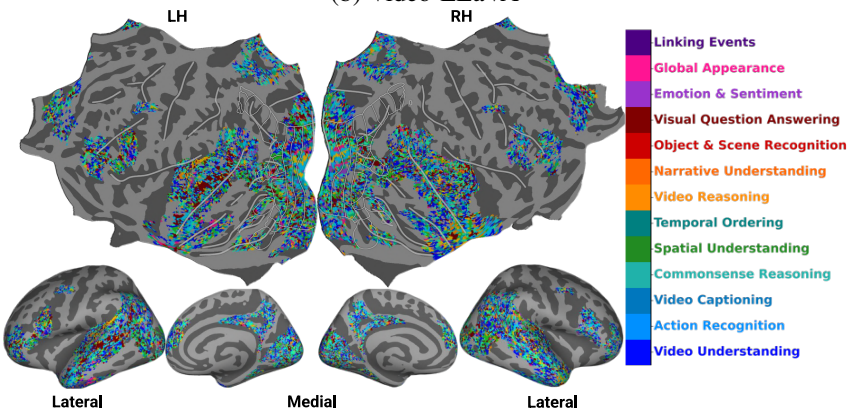
We also perform quantitative analysis to show a layerwise normalized brain alignment across three models, as shown in Fig. 25. Instruction prompts (Qwen-2.5-VL-7B; green solid) show a clear hierarchy: alignment rises from early layers and is strongest in mid to late layers. A non-language control (same model, orange dashed) flattens toward early layers, indicating that the hierarchy depends on instruction semantics rather than generic prompting. The smaller 3B model (light-green squares) exhibits the same shape with lower amplitude, demonstrating scale robustness of the hierarchy. Overall, we observe a significant shift in preferred layer under instruction vs. non-language, where a positive layer-trend is only observed for valid natural language instructions.



(a) InstructBLIPVideo

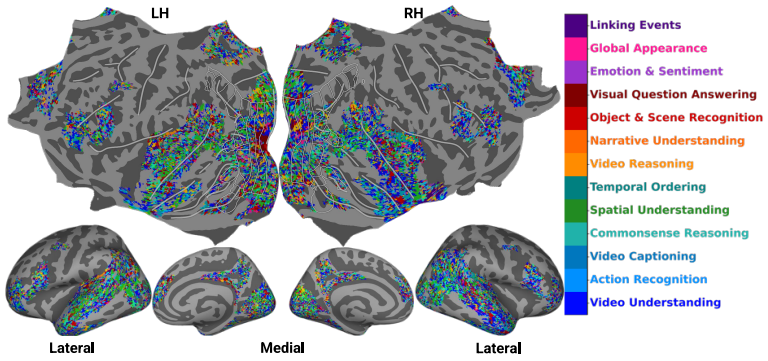


(b) Video-LLaVA

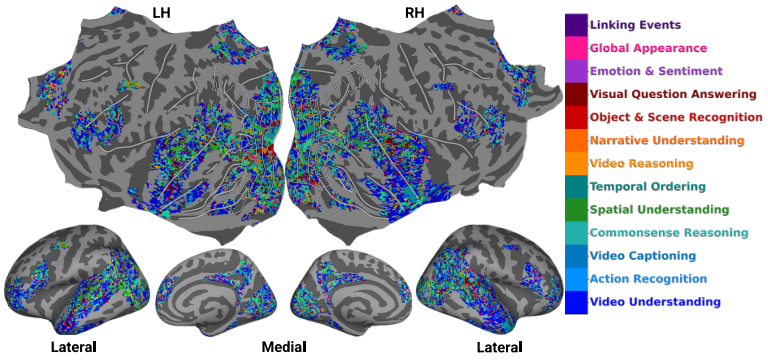


(c) LLaVA-NeXT-Video

Figure 19: Each voxel is color coded with the instruction (out of 13) that led to the highest normalized brain alignment. The color bar highlights color codes for each instruction. The voxels are projected onto the flattened cortical surface averaged across all 6 subjects for 3 video MLLM (InstructBLIPVideo, Video-LLaVA and LLaVA-NeXT-Video).



(a) LLaVA-OneVision



(b) VideoChat-R1

Figure 20: Each voxel is color coded with the instruction (out of 13) that led to the highest normalized brain alignment. The color bar highlights color codes for each instruction. The voxels are projected onto the flattened cortical surface averaged across all 6 subjects for 2 video MLLM (LLaVA-OneVision, VideoChat-R1).

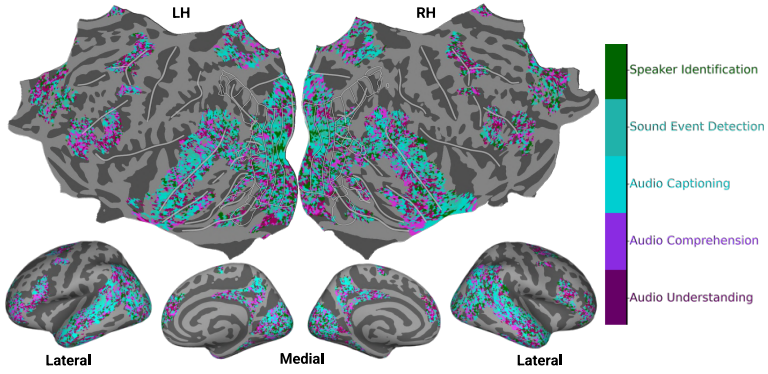
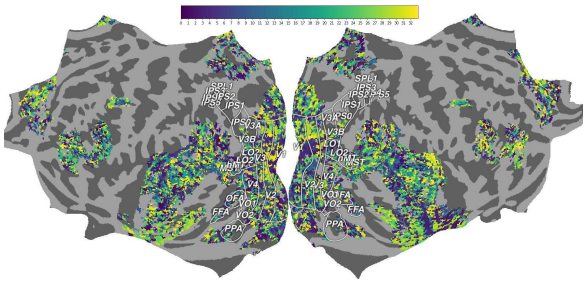
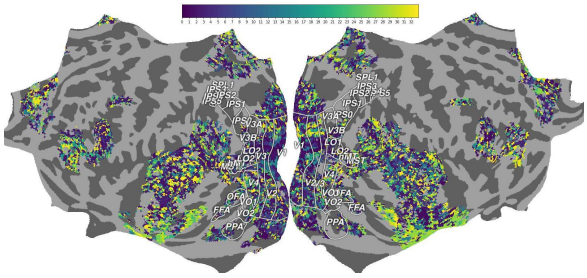


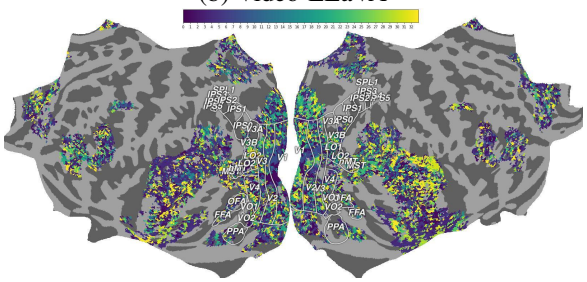
Figure 21: Kimi-Audio: Each voxel is color-coded with the instruction (out of 5) that led to the highest normalized brain alignment. The color bar highlights color codes for each instruction. The voxels are projected onto the flattened cortical surface of average across subjects on 'fsaverage' surface.



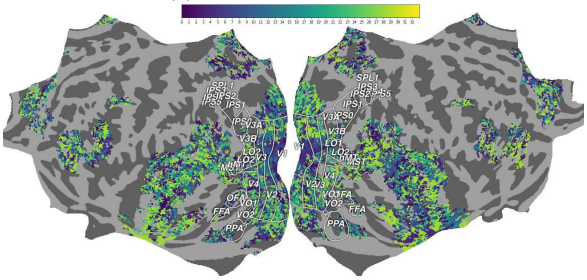
(a) InstructBLIPVideo



(b) Video-LLaVA

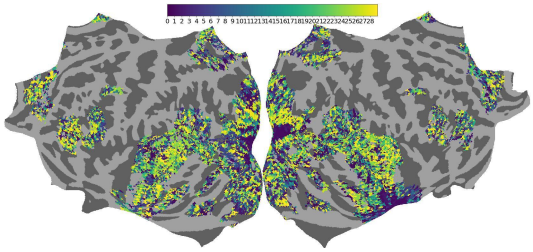


(c) LLaVa-NeXT-Video



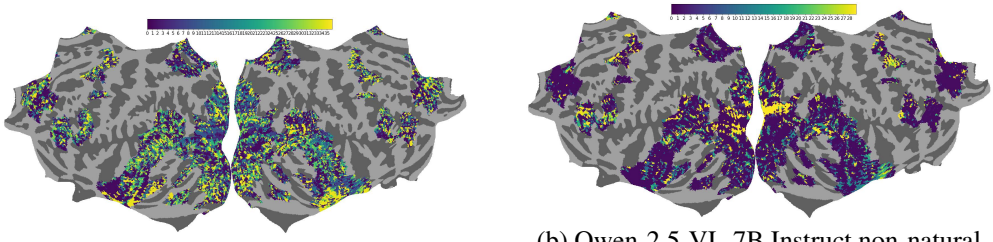
(d) LLaVA-OneVision

Figure 22: Each voxel is color coded with the video MLLM layer number (out of 33) that led to the highest normalized brain alignment. The color bar highlights color codes for each layer. The voxels are projected onto the flattened cortical surface of average across all 6 subjects on 'fsaverage' surface for four MLLMs.



(a) Qwen-2.5-VL-Audio

Figure 23: Each voxel is color coded with the video MLLM layer number (out of 29) that led to the highest normalized brain alignment. The color bar highlights color codes for each layer. The voxels are projected onto the flattened cortical surface of average across all 6 subjects on 'fsaverage' surface for four MLLMs.



(a) Qwen-2.5-VL-3B Instruct

(b) Qwen-2.5-VL-7B Instruct non-natural language prompt

Figure 24: (a) Qwen-2.5-VL-3B Instruct and (b) Qwen-2.5-VL-7B Instruct non-natural language prompt (layer-wise alignment): Each voxel is color coded with the MLLM layer number (out of 29) that led to the highest normalized brain alignment. The color bar highlights color codes for each layer. The voxels are projected onto the flattened cortical surface of average across subjects on 'fsaverage' surface.

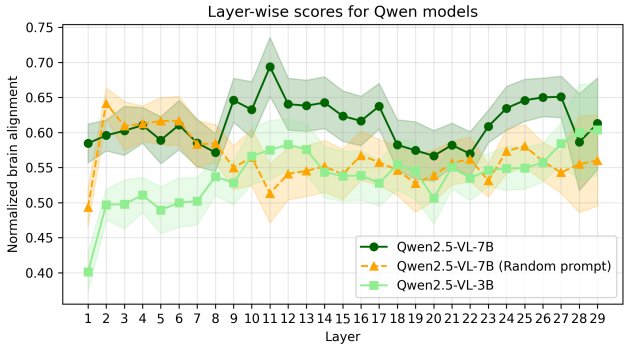


Figure 25: Layerwise normalized brain alignment for Qwen models.

N Self-Controlled Experiments: Comparing Models Before and After Instruction Tuning

Comparing the same backbone before and after instruction tuning is the most direct way to reduce architectural confounds when assessing differences in brain alignment. Similar to InternVL-8B, we use same Qwen-2.5 family and the same scale (7B): Qwen-2.5-VL-7B-Instruct and Qwen-2.5-VL-7B models, and Qwen-2 family and the same scale (7B): Qwen-2-VL-7B-Instruct and Qwen-2-VL-7B models, so that both models receive the same visual input. This keeps architecture family, parameter size, and modality comparable, and highlights the effect of instruction tuning. For both models, we issue the same instruction ("Describe the video") and extract the resulting instruction-specific representations across layers. This analysis shows how instruction tuning on the same model category changes layer-wise representations and improves brain alignment.

Qwen-2.5-VL: before vs. after instruction tuning: Using brain predictions across layers for Qwen-2.5-VL-7B-Instruct and Qwen-2.5-VL-7B models, we compute per-layer Δ alignment as

Δ alignment per layer = Normalized brain alignment (Qwen-2.5-VL-7B-Instruct) - Normalized Brain alignment (Qwen-2.5-VL-7B)

We make the following observations: (i) All layers improve (all Δ >0), (ii) Depth trend: Δ decreases with depth (early > mid > late): early 0.177, mid 0.156, late 0.139; Spearman ρ = -0.77, p $\ll$ 0.001. (ii) The center-of-mass of the improvement is near Layer 14/29 (normalized 0.47). Thus, while the absolute alignment peak under instruction occurs in mid to late layers (Fig. 26), the incremental benefit over the pretrained baseline is strongest early and remains positive throughout depth.

Table. 16 shows quantitative analysis of Δ alignment across families, we make following observations: (i) instruction tuning preserves the same alignment hierarchy (mid→late layers peak) and shifts preferred processing toward later layers (positive shift in preferred layer), (ii) the locus of the gain is family-specific: Qwen shows larger early-layer increases, whereas InternVL shows mid/late-layer increases—consistent with non-uniform representational shifts rather than a uniform shift.

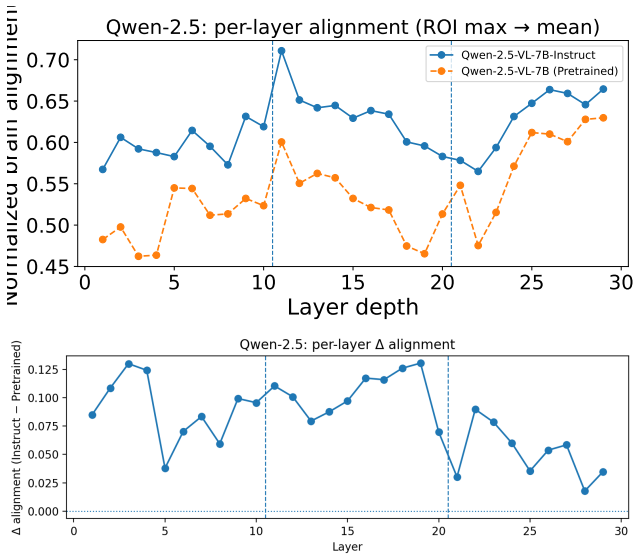


Figure 26: Qwen-2.5-VL: (top) Normalized brain alignment was computed before vs. after instruction tuning: Using brain predictions across layers for Qwen-2.5-VL-7B-Instruct and Qwen-2.5-VL-7B models. (bottom) Δ alignment per layer:= Normalized brain alignment (Qwen-2.5-VL-7B-Instruct) - Normalized Brain alignment (Qwen-2.5-VL-7B)

.

O Details of Semantic Task Group Analysis

To further examine how task-specific representations from instruction-tuned video MLLMs relate to regional brain predictivity, we group the 13 video tasks into five categories: perceptual visual

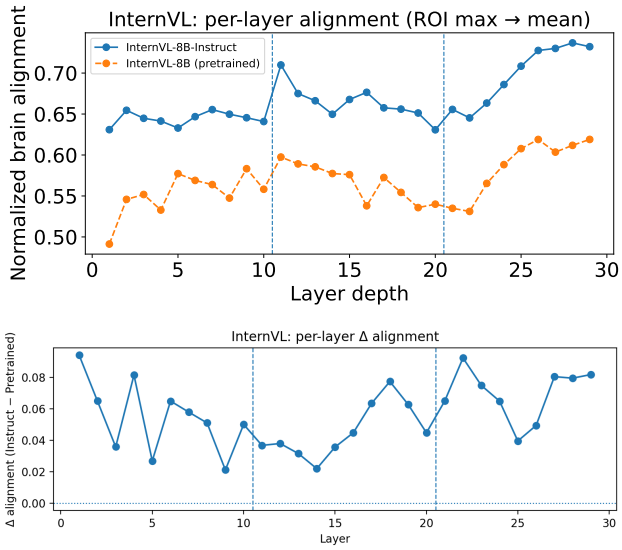


Figure 27: InternVL: Normalized brain alignment was computed before vs. after instruction tuning: Using brain predictions across layers for InternVL-8B-Instruct and InternVL-8B models.

Table 16: Matched-pair Δ alignment (IT − ICL). Early = layers 1–10, Mid = 11–20, Late = 21–29 (mean ± SD across layers). COM: center of mass of Δ across layers, normalized by depth. ρ: Spearman correlation between layer index and Δ.

Pair	Early Δ	Mid Δ	Late Δ	COM	ρ	p	Min Δ (layer)	Max Δ (layer)	Layers with Δ>0
Qwen-2.5-VL-7B	0.177 ± 0.021	0.152 ± 0.014	0.127 ± 0.025	0.474	-0.767	1.2×10^{-6}	0.088 (L29)	0.207 (L1)	29/29
Qwen-2-VL-7B	0.092 ± 0.024	0.070 ± 0.026	0.098 ± 0.025	0.503	$+0.110$	0.566	0.038	0.132	29/29
InternVL3-8B	0.092 ± 0.023	0.097 ± 0.019	0.112 ± 0.011	0.535	$+0.337$	0.074	0.056 (L5)	0.140 (L1)	29/29

processing, cognitive reasoning and integration, spatiotemporal understanding, high-level language and narrative understanding, and social and affective understanding. This grouping lets us compare which regions are best predicted by different types of task instructions (Fig. 28).

Spatiotemporal understanding. Temporal Ordering best predicts a larger share of voxels in the angular gyrus and posterior temporal lobe, whereas Spatial Understanding best predicts more voxels in dorsal parietal cortex (part of the dorsal visual pathway) and the anterior temporal lobe, consistent with the roles of these regions in event and spatial processing (Zacks et al., 2007; Baldassano et al., 2017). Spatial Understanding also best predicts more voxels in early visual areas, whereas Temporal Ordering best predicts more voxels in early auditory cortex, possibly reflecting sound-based cues to event boundaries (Belin et al., 2000). These preferences overlap substantially, consistent with the co-occurrence of spatial and temporal information in narrative and event-based processing.

P Details of Explained Variance Partitioning

To further quantify how IT-MLLMs capture shared and distinct neural processes across tasks, we use a variance partitioning approach. This analysis reveals the unique and overlapping contributions of individual task-specific representations to brain responses, enhancing our understanding of how brain regions differentially process multimodal information.

Variance partitioning. To disentangle task-specific instruction representations from multimodal instruction-tuned models, we used a variance partitioning approach (de Heer et al., 2017; LeBel et al., 2021). This method measures the overlap in brain variance explained by different task-specific instruction representations. Specifically, variance partitioning separates the brain response variance that can be attributed to two models based on their unique and overlapping contributions (Vaidya et al., 2022; Deniz et al., 2019). To perform this, for every pair of instruction representations, we fit separate encoding models for each space as well as a joint encoding model, obtained by concatenating the features. Using set arithmetic, we can then derive the size of the intersection $(NBA)_v^{1\cap 2}$=$(NBA)_v^1$+$(NBA)_v^2$-$(NBA)_v^{1\cup 2}$, where NBA refers to normalized brain alignment, v refers to a specific voxel, $(NBA)_v^1$ denotes alignment of model 1, $(NBA)_v^2$ denotes alignment of

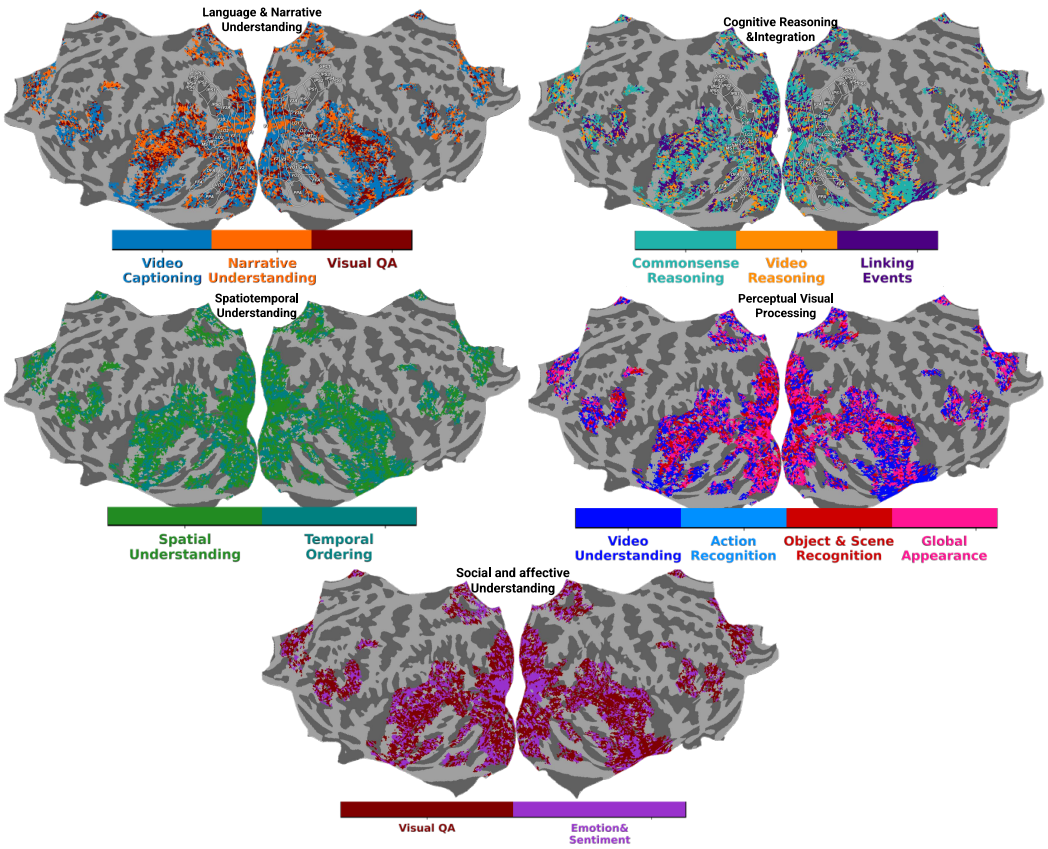


Figure 28: Semantic Task Group Analysis: Each voxel is colored by the task instruction yielding the highest normalized brain alignment. Voxels are projected onto the flattened cortical surface (averaged across subjects) for Qwen-2.5-VL.

model 2 and $(NBA)_v^{1\cup 2}$ denotes alignment of the joint model. Similarly, the unique contribution of model 1's feature space is computed as $(NBA)_v^{1\setminus 2}$=$(NBA)_v^1$-$(NBA)_v^{1\cap 2}$.

Partitioning explained shared and unique variance between task-specific instructions

While the previous analysis reveals that task-specific instructions from MLLMs modulate their representations based on distinct cognitive demands, we further examine the representations of task-specific instructions to measure the overlap in brain variance explained by MLLMs. We show analysis for all pairs from the 13 tasks in Table 17. Across nearly all task pairs, the whole brain region consistently exhibits the highest shared variance. Tasks that are conceptually or functionally related exhibit high shared variance in all regions, indicating similar cognitive processing demands. Higher-level semantic and reasoning tasks (e.g., Narrative Understanding, Commonsense Reasoning, Temporal Ordering) show increased unique variance in the language network, indicating language-specific processing distinct from visual features. High visual load tasks (e.g., Action Recognition, Object and Scene Recognition, Global Appearance) contribute more uniquely in visual cortex, especially when paired with non-visual tasks.

Shared and Unique Variance between Narrative Understanding and Remaining Task Instructions

Fig. 29 shows the shared variance of the Narrative Understanding task with other video tasks for Qwen-2.5-VL.

Table 17 presents shared and unique variance explained by pairs of video tasks using brain-informed models across three neural regions: whole brain, visual cortex, and language network. The results are averaged across subjects and show how well representations from each task pair align with brain activity in specific regions.

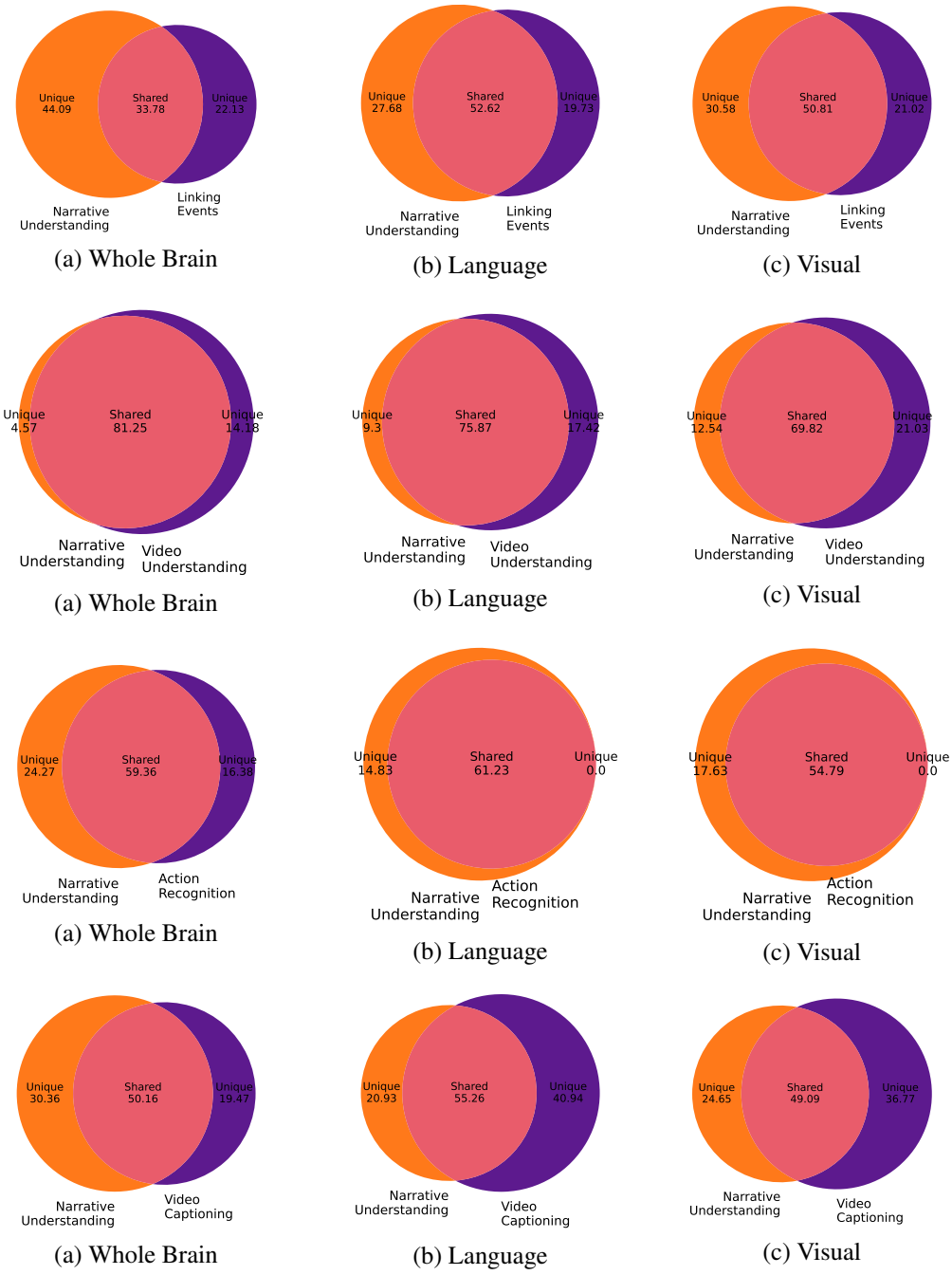


Figure 29: Shared and Unique Variance: Narrative Understanding vs. Linking Events Dark orange (left) shows variance unique to Narrative Understanding, indigo (right) shows variance unique to Linking Events, and the overlap indicates shared variance between both tasks.

Key Observations are as follows.

- Whole Brain Shows Dominant Shared Variance: Across nearly all task pairs, the whole brain region consistently exhibits the highest shared variance (often >80% in early task pairs). For example, the pair Action Recognition and Video Understanding (1–2) shows 90.69% shared variance, with very little unique variance from either task. This suggests high redundancy and common processing across tasks when considering global brain activity.

- Visual and Language Regions Yield More Balanced Partitioning: In contrast, visual and language-selective voxels exhibit lower shared variance and comparatively higher unique contributions from individual tasks. For the same task pair (1–2), shared variance in visual is 72.05%, and in language it is 77.46%, with higher unique components ($\sim$10-14%). This suggests that fine-grained processing differences are more pronounced in modality-specific regions.
- Task Similarity Reflects in Shared Variance: Tasks that are conceptually or functionally related (e.g., Narrative Understanding-Linking Events (10-13) or Emotion and Sentiment Analysis-Linking Events (11-13)) exhibit high shared variance in all regions, indicating similar cognitive processing demands. Conversely, task pairs with less conceptual overlap (e.g., Object Recognition-Commonsense Reasoning (5-6) or Visual QA-Object Recognition (3-5)) show lower shared variance and higher unique variance, especially in language and visual regions.
- Language Regions Show Selectivity for High-Level Tasks: Higher-level semantic and reasoning tasks (e.g., Narrative Understanding, Commonsense Reasoning, Temporal Ordering) show increased unique variance in the language network, indicating language-specific processing distinct from visual features. For instance, pair 6-13 (Commonsense Reasoning-Linking Events) yields 16.75% unique variance for Linking Events in the language network.
- Visual Cortex Captures Scene and Action Differentiation: Tasks with high visual load (e.g., Action Recognition, Object and Scene Recognition, Global Appearance) contribute more uniquely in the visual cortex, especially when paired with non-visual tasks.

		Whole Brain			Visual			Language		
Task1	Task2	Shared	Uniq1	Uniq2	Shared	Uniq1	Uniq2	Shared	Uniq1	Uniq2
1	2	90.69	5.26	4.05	72.05	13.91	14.04	77.46	12.07	10.47
1	3	83.53	10.05	6.42	73.67	10.28	16.05	77.05	10.72	12.23
1	4	84.51	9.65	5.84	71.87	13.82	14.31	75.97	12.27	11.76
1	5	79.16	13.51	7.33	66.82	14.35	18.83	73.47	13.07	13.46
1	6	81.48	13.34	5.18	68.44	17.28	14.28	73.59	15.37	11.04
1	7	83.07	10.44	6.49	71.99	11.88	16.13	75.20	12.30	12.50
1	8	81.25	14.18	4.57	69.82	17.63	12.54	75.87	14.83	9.30
1	9	86.94	7.57	5.50	73.42	10.25	16.34	78.27	9.05	12.68
1	10	84.55	9.06	6.39	73.46	10.59	15.95	76.42	10.32	13.26
1	11	85.44	8.51	6.05	74.92	11.12	13.96	76.56	10.96	12.48
1	12	82.46	11.66	5.88	72.88	12.75	14.37	76.02	12.50	11.48
1	13	91.81	4.20	3.99	74.92	11.82	13.26	80.06	10.00	9.94
2	3	83.59	9.72	6.69	73.14	11.39	15.47	74.15	12.80	13.05
2	4	86.25	7.40	6.36	73.32	13.52	13.16	74.41	12.14	13.45
2	5	77.09	14.33	8.58	64.55	17.14	18.31	70.20	15.08	14.72
2	6	79.86	13.99	6.15	69.43	17.86	12.71	73.10	14.96	11.94
2	7	83.62	9.46	6.92	72.53	12.65	14.82	71.61	14.43	13.95
2	8	81.30	13.10	5.60	67.98	18.96	13.05	72.05	16.07	11.88
2	9	86.64	7.42	5.93	73.55	12.35	14.11	75.55	10.62	13.83
2	10	85.25	7.97	6.78	72.98	12.28	14.73	73.28	12.51	14.21
2	11	84.70	8.31	7.00	73.27	12.25	14.48	72.48	13.27	14.25
2	12	82.97	11.16	5.88	73.06	14.41	12.54	72.99	14.99	12.02
2	13	91.78	3.66	4.55	74.89	12.59	12.52	78.19	9.77	12.03
3	4	68.68	13.67	17.64	68.53	18.38	13.09	71.98	14.19	13.83
3	5	50.07	24.61	25.32	52.60	24.08	23.32	60.68	17.79	21.53
3	6	61.39	21.67	16.94	61.59	22.97	15.44	65.21	18.68	16.12
3	7	65.21	17.99	16.80	64.73	20.33	14.94	66.85	17.80	15.35
3	8	66.30	20.20	13.49	61.04	23.96	15.00	62.43	21.86	15.71
3	9	70.23	13.71	16.06	70.07	16.68	13.25	72.20	12.52	15.28
3	10	66.99	13.00	20.01	68.60	15.97	15.42	64.43	15.79	19.78
3	11	68.07	14.39	17.54	66.84	17.50	15.66	66.97	16.85	16.18
3	12	61.81	19.24	18.95	65.81	19.69	14.50	67.09	17.92	14.99
3	13	83.92	6.44	9.64	71.83	16.87	11.31	76.76	12.86	10.38
4	5	55.03	24.36	20.61	53.05	20.94	26.00	59.06	18.82	22.13
4	6	61.72	25.66	12.62	59.66	24.72	15.62	63.75	21.99	14.26
4	7	69.00	17.62	13.38	66.08	17.45	16.47	67.89	17.50	14.61
4	8	63.88	21.85	14.27	60.24	23.59	16.17	65.25	19.95	14.80
4	9	71.16	16.55	12.28	65.51	18.15	16.34	68.66	16.14	15.19
4	10	66.37	18.11	15.53	63.85	17.11	19.04	57.73	20.94	21.33
4	11	72.37	13.56	14.07	70.00	13.01	16.99	70.64	13.35	16.02
4	12	66.38	18.76	14.86	64.80	17.67	17.53	67.94	17.21	14.85
4	13	86.69	6.09	7.23	71.23	16.28	12.49	76.56	13.87	9.57
5	6	50.13	27.24	22.63	51.63	27.81	20.56	58.56	23.05	18.39
5	7	49.08	24.63	26.29	53.55	25.15	21.30	55.77	24.66	19.57
5	8	47.03	27.55	25.43	53.22	28.86	17.93	53.88	26.92	19.21
5	9	55.06	21.61	23.34	56.84	24.75	18.42	62.62	19.24	18.15
5	10	47.76	23.54	28.70	55.84	22.99	21.17	54.52	22.48	23.00
5	11	52.17	22.58	25.25	57.44	22.32	20.24	57.94	22.48	19.58
5	12	47.50	26.51	25.99	56.38	25.48	18.15	58.21	23.50	18.29
5	13	79.36	6.98	13.67	66.31	16.96	16.74	71.80	12.91	15.29
6	7	60.01	17.04	22.96	59.05	17.09	23.86	61.14	18.01	20.84
6	8	54.31	21.48	24.22	57.44	21.55	21.01	62.62	18.13	19.25
6	9	64.33	13.06	22.61	60.10	16.20	23.69	64.68	13.72	21.60
6	10	57.84	16.91	25.25	61.41	14.59	24.00	61.01	16.15	22.84
6	11	62.94	14.26	22.81	62.17	15.15	22.68	63.32	15.40	21.28
6	12	55.82	19.64	24.54	60.18	17.37	22.45	60.36	18.93	20.71
6	13	81.42	5.21	13.37	67.46	13.51	19.02	71.93	11.31	16.75
7	8	58.19	23.15	18.65	60.58	23.47	15.95	61.00	20.86	18.13
7	9	70.87	14.02	15.11	70.43	15.05	14.51	71.25	12.70	16.05
7	10	68.57	12.51	18.92	67.67	13.27	19.06	63.76	14.39	21.84
7	11	60.77	18.94	20.29	58.79	21.23	19.98	55.14	21.77	23.09
7	12	66.57	17.86	15.57	67.97	17.05	14.98	67.18	17.38	15.44
7	13	85.27	6.01	8.72	72.66	15.56	11.78	74.88	13.08	12.03
8	9	62.84	15.99	21.18	63.11	15.66	21.22	68.03	13.67	18.31
8	10	60.10	17.38	22.52	59.39	16.80	23.81	60.46	16.80	22.74
8	11	60.31	14.63	25.07	61.67	13.24	25.09	61.38	15.64	22.98
8	12	60.04	18.69	21.28	62.31	17.41	20.28	65.74	16.70	17.56
8	13	81.06	5.66	13.27	68.01	14.38	17.61	74.50	11.65	13.85
9	10	69.21	14.34	16.44	68.83	12.98	18.19	67.69	15.88	16.44
9	11	70.80	13.15	16.05	69.96	14.08	15.96	70.82	14.04	15.15
9	12	69.68	16.60	13.72	70.09	14.45	15.46	70.62	16.10	13.29
9	13	87.40	5.23	7.37	72.02	15.46	12.53	77.48	12.70	9.82
10	11	68.63	16.35	15.02	67.96	16.43	15.61	64.85	19.12	16.04
10	12	65.06	20.66	14.27	63.79	21.85	14.36	61.84	23.65	14.50
10	13	85.63	6.39	7.99	72.34	16.92	10.73	75.85	14.09	10.06
11	12	61.95	22.51	15.54	65.60	19.55	14.85	63.80	21.51	14.69
11	13	86.42	6.00	7.58	74.60	14.29	11.11	76.83	12.89	10.28
12	13	83.82	5.77	10.41	71.56	15.38	13.06	75.37	12.20	12.43

Table 17: Variance partitioning for all the 13 video tasks averaged across all subjects for whole brain, visual and language regions with Qwen-2.5-VL model. Tasks are as follows: (1) Action Recognition (2) Video Understanding (3) Visual Question Answering (4) Video Captioning (5) Object and Scene Recognition (6) Commonsense Reasoning (7) Spatial Understanding (8) Temporal Ordering (9) Video reasoning (10) Narrative Understanding (11) Emotion and Sentiment Analysis (12) Global Appearance (13) Linking Events.

Q Extended Discussion

(1) ICL-MLLM representations are strongly coupled to instruction-text semantics, whereas IT-MLLM representations show weak coupling to instruction-text semantics, suggesting that IT-MLLMs produce representations organized by what the task is rather than by surface instruction wording.

(2) Although instruction-tuned video MLLMs show strong brain alignment across the whole brain (including language, visual, and auditory regions), audio MLLMs show stronger alignment primarily in auditory and language-related areas such as the middle frontal gyrus (MFG). This suggests that instruction-tuned audio MLLMs capture features relevant to auditory processing, consistent with patterns reported in previous studies (Oota et al., 2024a; 2025b). However, their performance remains comparable to non-instruction-tuned multimodal models, indicating that further improvements are needed for instruction-tuned audio MLLMs to better capture brain-relevant representations, an effort that aligns with recent work on inducing brain-relevant biases in model design (Moussa et al., 2025; Vattikonda et al., 2025).

(3) Task-specific instructions are effective at predicting multimodal brain activity across different regions, indicating that both IT video and audio MLLMs produce distinct task-conditioned representations. These representations support region-differentiated brain alignment patterns, unlike prior work by Oota et al. (2025a), which did not observe such differentiation when using unimodal stimuli (e.g., static images). Specifically, certain audio instructions, such as audio captioning and audio understanding, show stronger alignment with language-related regions, while tasks such as sound event detection align more with the auditory cortex and temporal lobe. These findings indicate that IT-MLLMs can serve as a useful framework for systematically varying task goals through instructions while holding the stimulus fixed, allowing researchers to examine task-conditioned brain alignment patterns.

(4) The observed correspondence between IT-MLLM layers and the brain's processing hierarchy is consistent with structured representational patterns, with shallow layers aligning more with sensory regions and deeper layers aligning more with higher-order semantic regions. This layered alignment improves interpretability and supports the use of IT-MLLMs as probes for studying how task-driven information is differentially encoded across cortical regions.

Our findings also indicate that despite the growing popularity of instruction-tuned video and audio MLLMs for handling generic task instructions, we are still far from fully understanding how language-based instructions modulate information flow through model layers and how fine-grained details are differentially represented across layers. Future work should focus on leveraging the alignment strengths of these models using more fine-grained instruction-driven prompts, similar to controlled stimulus paradigms in neuroscience, to deepen our understanding of how task-conditioned representations relate to neural responses.

R Discussion on Controlling Architectural and Pretraining Differences Across MLLMs

MLLMs differ in pretraining corpus and design. We therefore took three complementary steps to ensure fair and scientifically valid comparisons.

- (i) Controls in the evaluation protocol: Across all models, our encoding pipeline is the same as we use the same instruction templates, same video and audio datasets as input for feature extraction, use of the same cross-subject prediction scores for estimating normalized brain alignment per model.
- (ii) Matching scale: While all the MLLMs we tested are 7B to 8B models across both instruction-tuned video and audio MLLMs and in-context learning pretrained MLLMs. Thus, model training sizes close to several million across all instruction-tuned video MLLMs, with varying modality mixes and task diversity.
- (iii) Acknowledging training-data heterogeneity and checking robustness: The instruction-tuned video MLLMs differ substantially in training sources: Some (like LLaVA-NeXT-Video and Video-LLaVA) include explicit video instruction datasets, while others (like

InstructBLIPVideo and LLaVA-OneVision) primarily adapt image-based instructions to video via frame sampling. However, across all instruction-tuned videos MLLMs show similar normalized brain alignment to our results suggesting that depth alone does not account for the observed performance differences.

We also acknowledge that testing a larger number of models within a given class can help determine if the effect of modality is robust across various model configurations.

Literature precedent: Our approach follows established practice in neuro-AI, where models with differing training corpora/architectures are compared under a shared brain-encoding protocol and ceiling normalization to study representational alignment (Schrimpf et al., 2021; Toneva & Wehbe, 2019; Antonello et al., 2021; Aw & Toneva, 2023; Loong Aw et al., 2024). Specifically, the extensive precedent in the literature, from studies comparing 43 models (Schrimpf et al., 2021) to those examining 101 models (Antonello et al., 2021) in language models. Similarly Oota et al. (2024a) and Antonello et al. (2024) compared several text and speech models during language comprehension. These studies demonstrate that this approach is both valid and valuable for understanding the relationship between artificial and biological language processing. It is important to observe that all the above studies utilize a number of models that are different in training architecture and training datasets, however the primary goal of all these studies is to investigate how close the semantic representations captured by each model aligns with brain-relevant semantics. We have added pretraining corpus details across MLLMs in Table 18.

Table 18: **Training data and scale of compared multimodal large language models (MLLMs).**

Model	Training Data	Size
InstructBLIP-Video	26 public datasets: MSCOCO, TextCaps, NoCaps, VQA v2, iVQA, MSRVTT-QA, MSVD-QA	Not specified
Video-LLaVA	LAION-CC-SBU (558K), Valley (702K), LLaVA-Instruct (665K image-text + 100K video-text)	$\sim$1.3M total
LLaVA-NeXT-Video	LLaVA-Video-178K synthetic + real video QA/caption data	$\sim$1.4M total
Qwen-2.5-VL	Web-scale 4T tokens (text + multimodal), post-trained on 2M mixed samples	4T tokens + 2M samples
VideoChat-R1	Qwen-2.5-VL backbone, 18K video instruction samples (temporal grounding, tracking, QA)	18K samples
LLaVA-One-Vision	OneVision dataset (3.2M single-image + 1.6M multi-image/video), Evo-Instruct (143K)	$\sim$5M total
TVLT	HowTo100M, YTTemporal180M (video-audio pairs)	Large-scale (not stated)
VideoMAE	Kinetics-400, Something-Something v2, Epic Kitchens 100	$\sim$550K videos
AST	AudioSet, ESC-50, Speech Commands	$\sim$2+M audio clips

S Normalized brain alignment: cross-subject vs. repeat-based EV ceiling

We also compute the explainable variance (EV) using repeated test movies and perform thresholding on EV voxels. The EV is computed now based on (Schoppe et al., 2016). Using EV with threshold of 0.05, the normalized brain alignment on this mask. For fair comparison, we used the same reliability threshold (≥ 0.05) as in the cross-subject ceiling analysis.

From the Table 19, we find that the normalized brain alignment computed with the repeat-based EV ceiling is very similar to that obtained with the cross-subject ceiling; the model ranking is unchanged. We also include cortical flatmaps of repeat-based explainable variance (EV) for each participant in the Fig. 30, showing the spatial distribution of reliable voxels (EV ≥ 0.05).

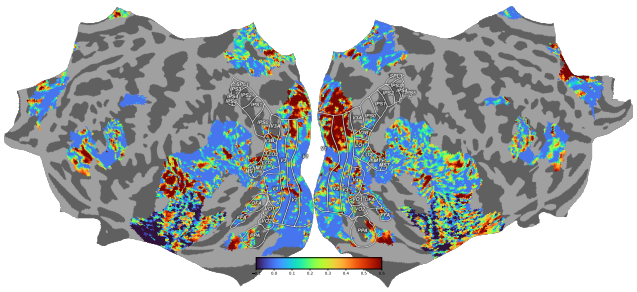
Table 19: Normalized brain alignment: cross-subject vs. repeat-based EV ceiling

Model	Cross-subject	Repeat EV	Δ (Repeat $-$ Cross)	$\%\Delta$
InstructBLIP Video	0.669	0.645	-0.024	-3.64%
Video-LLaVA	0.650	0.652	$+0.002$	$+0.35\%$
LLaVA-NeXT-Video	0.678	0.642	-0.037	-5.39%
Qwen-2.5-VL-7B	0.746	0.689	-0.056	-7.54%
LLaVA-OneVision	0.666	0.639	-0.028	-4.13%

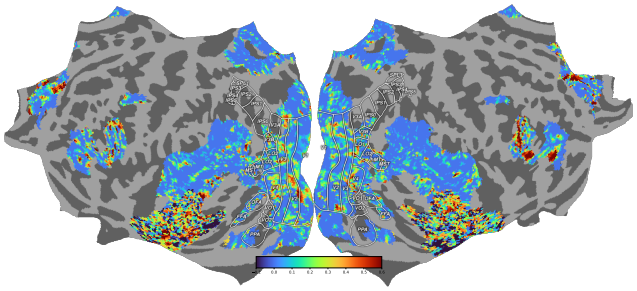
T Cross-Subject Generalization Analysis

To address potential concerns about subject-level generalization, we conducted a held-out-subject evaluation: encoding models trained on one subject were evaluated on a held-out subject. Results across all train-test subject pairs are reported in Table 20.

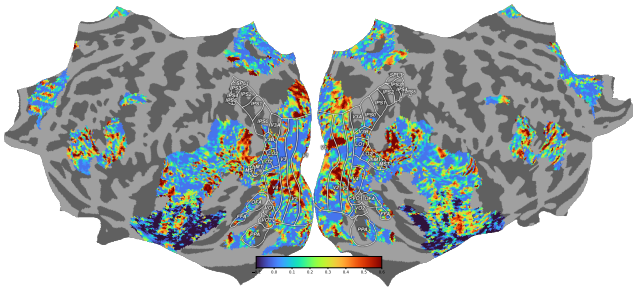
Comparison to within-subject performance.



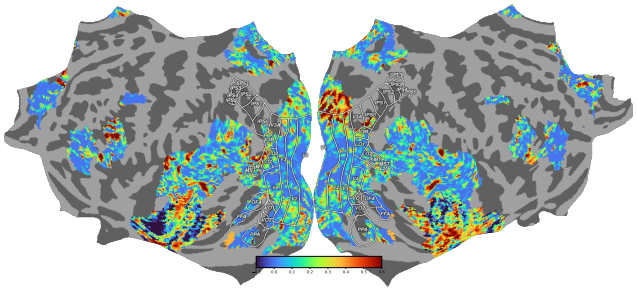
(a) Subject-01



(a) Subject-02



(b) Subject-03



(c) Subject-05

Figure 30: Estimated explainable variance for all six participants for the naturalistic movie watching. Explainable variance scores for each voxel in each subject are projected onto the subject's flattened cortical surface.

Setting	Mid Layers (L14-15)	Late Layers (L27-29)	Cross/Within Ratio
Within-subject	0.661	0.734	–
Cross-subject	0.669	0.757	1.01× / 1.03×

Methodological note. Because stimulus features are identical across subjects while neural responses are subject-specific, naive multi-subject concatenation would duplicate inputs with different targets, confounding interpretation. We therefore use train-on-one-subject, test-on-another-subject transfer as a cleaner test of cross-subject generalization.

Table 20: Cross-subject generalization. Encoding models trained on one subject and evaluated on a different held-out subject. Cross-subject alignment is comparable to within-subject performance (Mid: 1.01×; Late: 1.03×), indicating that deeper task-conditioned representations transfer across participants.

Train→Test	L14	L15	L27	L28	L29	Mid Mean	Late Mean
S1→S2	0.502	0.668	0.712	0.705	0.698	0.585	0.705
S1→S3	0.484	0.679	0.781	0.770	0.761	0.581	0.770
S1→S5	0.638	0.646	0.792	0.684	0.790	0.642	0.755
S3→S1	0.749	0.721	0.806	0.848	0.852	0.735	0.835
S3→S2	0.723	0.702	0.764	0.792	0.781	0.713	0.779
S3→S5	0.640	0.737	0.784	0.892	0.911	0.689	0.862
S5→S1	0.732	0.654	0.832	0.871	0.842	0.693	0.848
S5→S2	0.711	0.661	0.792	0.821	0.806	0.686	0.806
S5→S3	0.707	0.619	0.762	0.894	0.845	0.663	0.834
Average	**0.654**	**0.676**	**0.781**	**0.808**	**0.809**	**0.669**	**0.757**

U LIMITATIONS

One possible limitation of our study lies in interpreting the differences in brain alignment between instruction-tuned video and audio MLLMs. The models we evaluate differ in several aspects, including the amount of training data and the specific objective functions used during training. To address this concern, we evaluated multiple models of each type, spanning a range of training objectives and dataset sizes, and found that our key results generalize within both video and audio MLLM categories. Still, it is possible that some of the differences in brain alignment may still be influenced by confounding factors related to model architecture, training objectives, or data scale. Future work should explore these questions using models that are more tightly controlled across these dimensions. A second limitation is that participants passively watched the movies and did not perform the tasks specified in the model instructions. Our results therefore show that varying instructions on the model side produces representations with different brain-predictivity patterns for the same naturalistic responses; they do not demonstrate instruction-induced changes in participants' neural representations or that individual ROIs perform the prompted tasks. Directly testing correspondence between task conditioning in MLLMs and task-dependent cognition in humans will require neuroimaging datasets in which participants perform multiple instructed tasks on matched audiovisual stimuli. Even in the matched Qwen and InternVL pretrained–instruction-tuned comparisons, the instruction-tuned checkpoints differ from their pretrained counterparts in both the post-training objective and exposure to additional training data. These comparisons therefore estimate the effect of the released base-to-instruction-tuned transition while controlling for architecture, model scale, and pretrained initialization, but they cannot isolate the instruction objective from the accompanying post-training data. A decisive causal test would require an otherwise identical checkpoint trained on the same additional data using a non-instruction objective, which is not publicly available for the evaluated model families.

V MODEL GENERATED OUTPUTS ACROSS INSTRUCTIONS

Tables 21, 22, 23, 24, 25 and 26 show model generated outputs for a sample video from the Movie10 dataset using InstructBLIPVideo, Qwen-2.5-VL, Video-LLaVA, LLaVa-NeXT-Video, LLaVA-OneVision and VideoChat-R1 models, respectively. Similarly, Tables 27 and 28 show model generated outputs for a sample video from the Movie10 dataset using Qwen-2.5 Audio and Kimi-Audio models.

Table 21: Outputs from InstructBLIPVideo for a sample clip from the Movie10 dataset.



Frame 1 | Frame 3 | Frame 8 | Frame 13 | Frame 16

Instruction	**Model Output**
Action Recognition	The main event captured in the video is a man sitting at a bar with a glass of beer, engaging in conversation with another man. The man is wearing glasses and is sitting in a dining area, surrounded by other people, possibly in a bar or restaurant. The scene also features a man wearing a towel, which is wrapped around his neck.
Video Understanding	In the scene, a man is sitting in a bar with a beer in his hand, engrossed in a conversation with another man. The two men are sitting at a table, with one of them holding a glass of wine. There are multiple bottles of beer visible in the background, placed on a table and around the bar area. A few other people can also be seen in the scene, possibly patrons or other bar visitors, scattered around the bar and interacting with the two main characters. The scene is set in a dimly lit environment with some lighting, providing a relaxed atmosphere for the men's conversation.
Visual Question Answering	There are two people in the scene, one man and one woman. They are having a conversation at a bar, sitting at a table, and drinking.
Video Captioning	A man with glasses sitting at a table talking with another man.
Object and Scene Recognition	In the scene, there is a man wearing glasses sitting in a bar, talking to another person, who is also sitting in a chair next to him. The man is holding a glass of beer, while the other person is holding a cup.
Commonsense Reasoning	The man is sitting in a bar with a glass of wine, engaging in conversation with another man, possibly discussing something. This action could have motivated the man to take this step for various reasons. It could be that he is seeking to socialize, share his thoughts or ideas, or simply engage in casual conversation with a friend or acquaintance.
Spatial Understanding	Restaurant
Temporal Ordering	The scene depicted in the video shows a man sitting at a bar, engaging in a conversation with another man. The man is wearing glasses and is holding a glass in his hand, which suggests that he is drinking or holding a drink. They appear to be having a casual conversation, possibly discussing something of interest or sharing a story.
Video reasoning	In this video, there is an unusual scene where a man with glasses is having a conversation with another man in a bar, and they are both holding glasses. This is an unusual scene because it is not a typical social setting for two men to be having a conversation in a bar. They are sitting in a dark, dimly lit environment, which adds to the unusual nature of their interaction.
Narrative Understanding	The central conflict in the movie is the man sitting in a bar talking to another man, who is wearing glasses. The man in the bar is having a conversation with another man, possibly discussing something of interest. It is not specified what the topic of their conversation is, but it is likely related to the movie's storyline.
Emotion and Sentiment Analysis	The video shows a man sitting at a bar with a friend, talking and having a conversation, which suggests a social and relaxed atmosphere. The man is wearing glasses, and the scene is set in a dimly lit bar, which creates a more intimate and cozy atmosphere.
Global Appearance	Man with glasses and a towel
Linking Events	The scene depicted in the image shows a man wearing glasses sitting at a bar, talking to another person, possibly a friend, while holding a glass of beer. This early event sets the stage for the later development of a story or a situation that may unfold in the video.

Table 22: Outputs from Qwen-2.5-VL for a sample video from the Movie10 dataset.



Frame 1 | Frame 3 | Frame 8 | Frame 13 | Frame 16

Instruction	Model Output
Action Recognition	The video appears to be a still from a movie or a scene from a film. It shows two individuals engaged in a conversation in what looks like a dimly lit bar or restaurant setting. The person on the left is wearing a suit and tie, holding a drink, and seems to be listening intently. The person on the right is gesturing with their hand, possibly emphasizing a point during the conversation. The atmosphere suggests a serious or intense discussion.
Video Understanding	The video depicts a scene from a movie or TV show, featuring two characters engaged in a conversation. The character on the left is dressed in a formal gray suit with a patterned tie and is holding a cigarette in his right hand. He appears to be listening intently to the other character, who is partially visible from behind. The setting seems to be an indoor environment, possibly a bar or a similar establishment, as suggested by the background elements such as a mirror and what looks like a bar counter.
Visual Question Answering	There are two people in the scene. One person is sitting at a table, wearing a suit with a patterned tie, and appears to be engaged in a conversation or interaction with another person who is partially visible from behind. The second person is facing away from the camera, and their attire includes a plaid shirt. The setting seems to be indoors, possibly in a bar or restaurant, given the background elements like the glassware and the dim lighting.
Video Captioning	A man in a suit is engaged in a conversation with another person in a dimly lit bar setting.
Object and Scene Recognition	In the video, there are two main individuals and several objects: **1. The Man in the Suit:** He is wearing a brown suit with a white shirt and a red paisley tie. He has dark hair styled neatly and is holding a glass of what appears to be a drink in his left hand. His expression seems serious or contemplative as he looks at the other person. **2. The Other Person:** Only the back of this individual's head is visible, suggesting they are facing away.
Commonsense Reasoning	In the video, the character on the left appears to be in a conversation with another person whose back is turned to the camera. The character's hand is extended towards the other person, possibly gesturing or pointing at something. This action could be motivated by several reasons: **1. Clarification or Explanation:** The character might be trying to explain something important or clarify a point in their conversation. **2. Emphasis:** The gesture could be used to emphasize a particular point they are making.
Spatial Understanding	The video appears to be from the movie "The Wolf of Wall Street", directed by Martin Scorsese and released in 2013. The setting looks like a bar or a lounge, which is consistent with the film's depiction of the characters' activities in the financial world. However, there are no specific landmarks visible in the image that would definitively identify the exact location within the movie. The scene is set in New York City, as the film is based on the real-life events.
Temporal Ordering	The image depicts a scene from a movie where two characters are engaged in a conversation. Here is a step-by-step description of the activity shown: **1. Setting:** The setting appears to be an indoor environment, possibly a bar or a restaurant, given the presence of a glass and what looks like a menu board in the background. **2. Characters:** - The character on the left is wearing a suit with a patterned tie and is holding a glass in his hand.
Video reasoning	The video appears to be from a scene in a movie or TV show, featuring two characters engaged in a conversation. The setting looks like a bar or a similar social environment, with dim lighting and a relaxed atmosphere. What might be considered unusual is the character on the right, who seems to be gesturing with their hand while holding a drink, which could indicate they are explaining something or making a point during the conversation. However, without additional context, it's difficult to determine if there is anything particularly.
Narrative Understanding	The movie depicted in the video is "The Wolf of Wall Street," directed by Martin Scorsese and starring Leonardo DiCaprio as Jordan Belfort. The film is based on the memoir of the same name by Belfort himself. **Main Storyline:** The story follows Jordan Belfort, a young stockbroker who rises to become one of the most successful and corrupt figures in the financial world during the late 1980s and early 1990s.
Emotion and Sentiment Analysis	The video depicts two individuals in what appears to be a dimly lit, possibly upscale setting, such as a bar or restaurant. The person on the left is dressed in a suit and tie, suggesting a formal or professional demeanor, while the person on the right is wearing a plaid shirt, which might indicate a more casual or relaxed attitude.
Global Appearance	In the video provided, there is no indication of significant changes in the characters' appearances or outfit changes. The man on the left appears to be wearing a brown suit with a patterned tie and a white shirt. He has dark hair styled neatly. The person on the right is seen from behind, wearing a plaid shirt with a blue and red pattern. There are no visible changes in their attire or appearance within this single frame.
Linking Events	In the video provided, we see a scene from the movie "The Wolf of Wall Street," where Leonardo DiCaprio's character, Jordan Belfort, is engaged in a conversation with another character. The setting appears to be a dimly lit bar or lounge, which is typical for scenes involving discussions about business deals or personal matters. An early event that could influence later developments in this scene might involve a previous interaction between Jordan and his current interlocutor.

Table 23: Outputs from Video-LLaVA for a sample clip from the Movie10 dataset.



Frame 1	Frame 3	Frame 8	Frame 13	Frame 16

Instruction	**Model Output**
Action Recognition	The main events in the video include a man sitting at a bar, holding a beer, and talking to another man. The man holding the beer is wearing a plaid shirt and glasses.
Video Understanding	The video features a man sitting at a bar with a drink in his hand. He appears to be enjoying the drink and is engaged in conversation with someone else at the bar. The man is wearing glasses and a plaid shirt, giving him a casual and relaxed demeanor. The bar is well-lit, with a variety of bottles and glasses on display, indicating a lively and bustling atmosphere. In the background, there are other patrons.
Visual Question Answering	There are three people in the scene. Two of them are sitting down, while the third man is standing up and holding a beer glass.
Video Captioning	A man in a blue and plaid shirt drinking beer at a bar.
Object and Scene Recognition	In the scene, the main objects visible are a man sitting at a bar with a beverage in a glass and another man sitting next to him. There are also two other men sitting at the bar, and a woman standing nearby. The bar is well-stocked with various bottles and cups. The man sitting at the bar is wearing glasses, a blue plaid shirt, and a beige blanket.
Commonsense Reasoning	The character in the image is holding a beer glass and appears to be in a bar or a similar setting. The motivation for the character to take this action could be due to various reasons, such as relaxation, socialization, or simply enjoying a beverage during leisure time. The character may have chosen to have a beer to unwind after a long day, to connect with others in a social setting, or simply to indulge in a favorite beverage.
Spatial Understanding	The video is taken from a bar, as indicated by the presence of multiple bottles on the bar and a man holding a beer glass. The video also shows the man standing in front of a bar, suggesting the location of the video is a bar or similar establishment.
Temporal Ordering	The video depicts a man sitting at a bar, holding a beer in his hand. He is wearing a plaid shirt and glasses. The scene is set in a dark room, and there are other people in the background. The man is seen drinking from the beer glass, which is placed on the bar counter. The camera captures the man's facial expressions and body language, highlighting his relaxed and social atmosphere.
Video reasoning	The unusual aspect of this video is that it features a man in a plaid shirt and glasses, who is sitting in a bar and drinking a beer. The man is also wearing a beige scarf, which adds a distinctive touch to his appearance. However, the most striking feature of this video is that the man is not only drinking a beer, but he is also holding a cell phone.
Narrative Understanding	The central conflict in the movie is a man's desire for a drink at a bar, but he is unable to get one due to a bartender's refusal. The man's frustration and disappointment lead to him making a scene and causing a commotion at the bar. The conflict is ultimately resolved when the man is given a drink, and he leaves the bar with a smile.
Emotion and Sentiment Analysis	In the video, the man in the blue shirt and the woman are seen laughing and smiling while enjoying drinks and conversation. The man in the plaid shirt is seen drinking beer and looking down, while the man in the blue shirt is holding a beer glass. The video conveys a sense of relaxation, camaraderie, and enjoyment among the characters.
Global Appearance	In the video, the main character, a man with a beard, is seen wearing a plaid shirt and a scarf. He is also seen wearing a different outfit in one of the scenes, which is not specified in the given information. The other man in the scene is wearing a dark shirt and glasses. There are also two other men in the background, one wearing a dark jacket and the other wearing a dark shirt.
Linking Events	The early event in the video is the man in the blue and white checkered shirt sitting at the bar with a beer in his hand. This scene introduces the setting and the man's presence at the bar, setting the stage for later developments. The man's interaction with other patrons, the woman, and the man in glasses adds depth and complexity to the narrative.

Table 24: Outputs from LLaVA-NeXT-Video for a sample clip from the Movie10 dataset.



Frame 1 | Frame 3 | Frame 8 | Frame 13 | Frame 16

Instruction	**Model Output**
Action Recognition	The main event captured in the video is a conversation between two men at a bar, where one of them is wearing a red and white checkered shirt and glasses, while the other man is wearing a dark suit and tie. They are discussing a topic, possibly related to a movie, as suggested by the context. There are also other patrons in the background, and a bartender is visible in the foreground.
Video Understanding	The video features a scene from the movie The Hangoverfeaturing actors Zach Galifianakis and Bradley Cooper. In the scene, Zach Galifianakis, portraying a character named Chip Baskin, is sitting at a bar with a group of people. He is seen holding a beer and appears to be in a state of disbelief as he looks on at the camera.
Visual Question Answering	There are several people in the scene. The man with glasses is sitting at a bar, holding a drink, and appears to be in a conversation with someone. Another man is standing at the bar, and there is a woman sitting nearby. There is a bartender standing at the bar, and a person in the background who is partially visible, but it's unclear what their role is in the scene.
Video Captioning	Two men are seen having a conversation in a bar, with the man in the plaid shirt appearing confused or disbelieving the man in the suit.
Object and Scene Recognition	In the scene, we can see a man in a suit sitting at a table with a woman and a bartender. There's also a man in a suit standing at a bar, a woman sitting at a table with a man, a woman standing at a bar, a man standing at a bar, and a man sitting at a table with a woman.
Commonsense Reasoning	In the video you provided, the character is seen holding up a beer bottle and making a face, which could be interpreted as a playful or humorous gesture, possibly to amuse others or to express a lighthearted or jovial moment. This action might be motivated by a desire to entertain, bond with others, or to simply enjoy a casual and relaxed moment.
Spatial Understanding	The video appears to have been taken from a scene in a movie, specifically "Goodfellas" featuring Robert De Niro. The scene is set in a bar, and the background includes a sign that says "The Goodfella's." This is a reference to the character Robert De Niro's character, who is a bartender in the movie. The specific setting and the sign are consistent with the movie's depiction of the bar.
Temporal Ordering	The video depicts two men sitting at a bar, engaged in a conversation. One man is wearing a suit, while the other is dressed in a plaid shirt. They are both holding drinks and appear to be having a casual conversation. The setting suggests they might be at a social event, such as a bar or a party, where people gather to relax and enjoy drinks.
Video reasoning	In the video, there is a scene where a man is seen sitting at a bar with another man. The man in glasses appears to be engaged in a conversation with someone off-camera, but the camera is focused on the other man, who is not the focus of the scene. The man with glasses is holding a beer and appears to be in the middle of a discussion, but his face is not shown in the shot.
Narrative Understanding	The movie "The Big Short" is a biographical comedy-drama film about the life of stockbroker Michael Burry, who predicts the financial crisis of 20008. The central conflict of the story is the housing bubble and the subprime mortgage market, which Burry sees as unsustainable and warns his colleagues about, but they dismiss his ideas as unrealistic.
Emotion and Sentiment Analysis	The characters in the video express a range of emotions, from surprise and shock to amusement and even a bit of confusion. The video itself is quite humorous, with the man in a suit and tie appearing to be in a state of disbelief or shock as the other man, who is wearing a plaid shirt and a hoodie, is seen drinking from his beer bottle.
Global Appearance	In the video, we can see a man with a beard who is wearing a suit at the beginning, but he is later seen in a more casual outfit with a plaid shirt. There are also other people in the scene, including a man with glasses and a red shirt, a man with a beard and a red shirt, and a woman in a red dress.
Linking Events	In the scene you've described, a man is shown sitting in a bar with a drink in front of him, looking at a TV. The TV shows a news segment about a man who has died, which is later revealed to be the man himself. This early event in the video influences the later developments because it sets up the narrative of the story, introducing the main character and establishing his presence in the scene.

Table 25: Outputs from LLaVA-OneVision Video for a sample clip from the Movie10 dataset.



Frame 1 | Frame 3 | Frame 8 | Frame 13 | Frame 16

Instruction	**Model Output**
Action Recognition	The video captures a conversation between two individuals at a bar, with one person gesturing and speaking while the other listens.
Video Understanding	The video takes place in a lively bar setting, creating a warm and inviting atmosphere. The background features a well-lit bar counter adorned with various bottles and glasses, and several people can be seen in the background, adding to the bustling ambiance. In the foreground, two individuals are seated at a table. One person is dressed in a gray suit with a red and white patterned tie, while the other is wearing a blue and red plaid shirt with a yellow jacket draped over their shoulders.
Visual Question Answering	There are two people in the scene, engaged in a conversation.
Video Captioning	A man with curly hair and glasses sits at a table in a bar, holding a glass of beer, and engages in a conversation.
Object and Scene Recognition	The main objects and people visible in the scene are two men sitting at a table in a dimly lit, bustling bar or restaurant. One man is dressed in a gray suit and tie, and the other is wearing a plaid shirt with a beige sweater over it, holding a drink. The background features a bar with various bottles and glasses, and other patrons are visible in the distance.
Commonsense Reasoning	The character took this action to emphasize their point in the conversation, which could have been motivated by the seriousness of the discussion or the need to convey a specific message.
Spatial Understanding	The video is taken from a movie scene, and it shows a bar.
Temporal Ordering	The video takes place in a dimly lit bar with a lively background filled with people and bottles on shelves. Two individuals are seated at a table, engaged in a conversation. The person on the left is dressed in a gray suit with a red and white patterned tie and is gesturing with their hand, while the person on the right, wearing a blue and red plaid shirt with a yellow sweater draped over their shoulders and holding a glass of beer, listens attentively.
Video reasoning	The unusual aspect is that the scene remains static with minimal changes in the individuals' expressions and the background, suggesting a continuous conversation or interaction.
Narrative Understanding	The central conflict involves two characters in a bar, with the character in the plaid shirt and yellow sweater being advised or persuaded by the other character. The conflict is resolved through a conversation, leading to a change in the character's demeanor and posture.
Emotion and Sentiment Analysis	The characters express a range of emotions, including surprise, concern, and contemplation. The video creates a sense of intrigue and emotional depth, drawing the viewer into the characters' conversation.
Global Appearance	The character on the right initially has curly hair and is wearing a blue plaid shirt with a beige sweater, but later, the character's hair appears straight and they are dressed in a dark suit with a red tie.
Linking Events	The man in the plaid shirt and glasses is seen holding a glass of beer and engaged in a conversation with the man in the gray suit, which sets the stage for the unfolding dialogue and interaction.

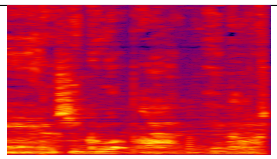
Table 26: Outputs from VideoChat-R1 for a sample clip from the Movie10 dataset.



Frame 1 | Frame 3 | Frame 8 | Frame 13 | Frame 16

Instruction	**Model Output**
Action Recognition	The video captures a scene from what appears to be a social setting, possibly a bar or restaurant. Two individuals are engaged in conversation. One person is wearing a plaid shirt and has a blanket draped over their shoulders, suggesting they might be cold or seeking comfort. The other individual is dressed in a suit and tie, indicating a more formal attire. The background shows other patrons and bar stools, contributing to the ambiance of a casual gathering.
Video Understanding	The video depicts a scene from what appears to be a movie or TV show, set in a dimly lit bar or lounge. Two individuals are seated at a table, engaged in conversation. The person on the left is wearing a dark suit and tie, while the individual on the right is dressed in a plaid shirt with a yellow scarf draped over their shoulders. Both appear to be holding drinks, suggesting they might be enjoying a casual evening out.
Visual Question Answering	There are two people in the scene. They appear to be engaged in a conversation at a bar or restaurant. One person is wearing a plaid shirt and has a blanket draped over their shoulders, while the other person is dressed in a suit. The setting suggests a casual social interaction.
Video Captioning	Two individuals engage in conversation at a bar, surrounded by dim lighting and patrons.
Object and Scene Recognition	The video shows two individuals seated at a bar, engaged in conversation. The person on the left is wearing a dark suit and appears to be listening attentively. The person on the right is wearing a plaid shirt with a yellowish blanket draped over their shoulders and is holding a drink. The background features a bar setting with various bottles and glasses, indicating a social or casual environment. The lighting suggests an indoor setting, likely during the evening.
Commonsense Reasoning	In the video, one character appears to be engaged in a conversation with another character at what looks like a bar or a similar social setting. The character on the right seems to be leaning forward slightly, which might indicate that they are listening intently or trying to emphasize a point during their conversation. The motivation for this action could be: 1. **Engagement**: The character might be trying to show interest or engagement in the conversation by leaning closer.
Spatial Understanding	The video appears to be from a movie or TV show set in a bar or restaurant. The setting includes a bar counter with bottles and glasses, suggesting it could be a scene from a film or series that takes place in a social or dining establishment. However, without additional context or recognizable landmarks, it's difficult to pinpoint the exact location or landmark shown in the video. It could be from a variety of films or shows that feature such settings.
Temporal Ordering	The video depicts two individuals seated at a bar, engaged in conversation. The person on the left is wearing a suit and appears to be listening attentively. The individual on the right is dressed casually in a plaid shirt with a blanket draped over their shoulders, suggesting they might be cold or seeking comfort. They both have drinks in front of them, indicating they are likely enjoying a social setting. The background shows other patrons and bar stools, contributing to the ambiance of a lively bar environment.
Video reasoning	The video appears to be from a scene in a movie or TV show, likely set in a bar or similar social setting. The unusual aspect here is the presence of a large, animated hand with a glove that resembles a cartoonish character, which seems out of place in a real-world setting like a bar. This suggests it might be a special effect or a prop used for comedic or dramatic purposes within the context of the film.
Narrative Understanding	The video appears to be from a scene in the movie "The Hangover," where the characters are at a bar, seemingly engaged in a conversation. The central conflict in this movie revolves around a group of friends who go on a bachelor party to Las Vegas but end up getting lost and having various mishaps.
Emotion and Sentiment Analysis	The video depicts two individuals engaged in what appears to be an intense conversation at a bar. The person on the right seems to be speaking with a sense of urgency or concern, as indicated by their open mouth and expressive eyes. The other individual is listening attentively, suggesting a moment of serious discussion. Overall, the scene conveys a sense of tension and engagement. The setting, with its dim lighting and crowded bar, adds to the atmosphere of a private yet public interaction.
Global Appearance	The video shows two characters sitting at a bar. The character on the left is wearing a dark suit and tie, while the character on the right is wearing a plaid shirt with a yellow blanket draped over their shoulders. There are no significant changes in their appearances or outfits throughout the video; they remain in these respective outfits for the duration of the frames provided.
Linking Events	In the video, two individuals are seated at a bar, engaged in conversation. The setting suggests a casual and possibly intimate interaction, which could imply that they are discussing something personal or significant. This early event of them sitting together and talking could influence later developments by establishing a foundation for their relationship or the context of their conversation. For instance, if this is part of a larger narrative, the initial interaction might set up themes such as friendship, trust, or conflict that will unfold in subsequent scenes.

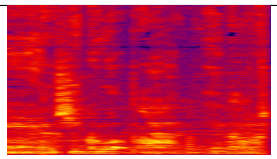
Table 27: Outputs from Qwen-2.5 Audio for a sample audio from the Movie10 dataset (Wolf of wallstreet).



Spectrogram

Instruction	**Model Output**
Audio Understanding	There's a male singing in English with an angry mood, saying 'And she grew up hot.
Audio Comprehension	They are talking and laughing.
Audio Captioning	The audio is about someone's upbringing.
Sound Event Detection	There are no background sounds or music, only speech. A male voice speaks in English with the phrase 'And she grew up hot.
Speaker Identification	A man speaking in English with an angry mood said 'And she grew up hot, you know?

Table 28: Outputs from Kimi-Audio for a sample audio from the Movie10 dataset (Wolf of wallstreet)



Spectrogram

Instruction	**Model Output**
Audio Understanding	The audio contains a speech where kids are talking by the door in English with a female voice and a neutral mood.
Audio Comprehension	The people are talking and there is music playing in the background.
Audio Captioning	A man is speaking loudly and aggressive.
Sound Event Detection	There is no speech in this audio.
Speaker Identification	The speaker in the audio is a Mandarin-speaking male, aged between 16-25 years old, with an angry mood.