Aligning MLLMs with the Latent Structure of Human Cognition via Behavior-Derived Semantic Dimensions

Ning E[1,2*] **Changde Du**[2,5*] **Yizhuo Lu**[2,3*] **Huiguang He**[2,3,4,1,5†]

[1]School of Biomedical Engineering, ShanghaiTech University, Shanghai, China
[2]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]School of Future Technology, University of Chinese Academy of Sciences, Beijing, China
[4]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[5]Zhongguancun Academy, Beijing, China

Abstract

Multimodal Large Language Models (MLLMs) have achieved impressive performance across a wide range of reasoning tasks, yet it remains unclear whether their internal representations and decision criteria align with the latent structure of human cognition. In this work, we propose a novel framework to align MLLMs with the multidimensional mental representations underlying human similarity judgements. We build on a previously established behavior-derived semantic embedding space, in which objects are represented along 66 interpretable dimensions derived from 4.70 million human Odd-One-Out (O1O) judgements. For each triplet, we infer the most salient decision dimension and convert it into a dimension-grounded linguistic rationale. This enables MLLMs to learn both the final choice and a justification consistent with the inferred behavioral criterion. Experimental results demonstrate that our approach improves model consistency with human similarity judgements while retaining competitive performance on standard multimodal benchmarks. Furthermore, using searchlight Representational Similarity Analysis (RSA) on an independent fMRI Natural Scenes Dataset (NSD), we observe increased representational alignment between model and human. Ultimately, our findings point to a data-driven route for incorporating human cognitive structure into MLLMs alignment. The code is available here.

1 Introduction

The advent of Multimodal Large Language Models (MLLMs) marks a major milestone in artificial intelligence, demonstrating strong performance across a spectrum of cognitive and reasoning tasks [Qwen Team, 2026, Bai et al., 2025a, Zeng et al., 2026] and naturally emerging human-like representations [Du et al., 2025]. Current alignment paradigms [Fu et al., 2023, Sundaram et al., 2024, Ouyang et al., 2022, Schulman et al., 2017, Rafailov et al., 2023] focus on outcome-level behavioral mimicry, optimizing models to generate outputs that match human preferences without deeper human cognitive logic. Relying on this shallow mimicry leaves it unclear whether models solve tasks using human-consistent decision criteria or by exploiting statistical shortcuts. To realize more robust generalization, intrinsic interpretability, and authentic anthropomorphism, MLLMs alignment requires moving beyond behavioral mimicry toward the latent structure of human cognition.

*Equal contribution.
†Corresponding author: huiguang.he@ia.ac.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

To probe this alignment gap, we utilize the Odd-One-Out (O1O) task, a cornerstone of cognitive psychology for evaluating conceptual representations and reasoning. In this task, a subject is presented with three objects and must identify the one that is least similar to the other two. Identifying the odd object requires accessing the latent semantic dimensions that humans use to structure their conceptual world. Evaluating MLLMs on this task demonstrates that, despite their robust general capabilities, a persistent performance gap remains in predicting human choices. This disparity reveals a misalignment: standard approaches primarily align models with raw behavioral outcomes (i.e., what humans choose), failing to capture the underlying human cognition that dictates the decision criteria (i.e., why they choose it).

Our key insight to bridge this gap is to transform the latent structure of human cognition into explicit and learnable supervision. To achieve this, we build on the previous work of Hebart et al. [Hebart et al., 2020, 2023], which established behavior-derived semantic embedding space. By analyzing 4.70 million human O1O judgements, they reveal interpretable 66-dimensional mental representations of natural objects underlying human similarity judgements. We adopt the 66-dimensional mental representations as the latent structure of human cognition to align MLLMs, enabling MLLMs to learn not only what humans choose, but also why they make choice. Concretely, for each triplet, we identify the most influential dimension driving human decision-making via the Jackknife [Mahner et al., 2025] method. However, a discrete semantic dimension is insufficient to bridge the gap with generative modality of MLLMs. To address this, we convert the discrete dimension into a dimension-grounded linguistic rationale, which serves as an explicit Chain-of-Thought (CoT) constraint.

Building on this formulation, we construct the alignment dataset to fine-tune MLLMs, thereby deeply aligning their reasoning processes with the latent structure of human cognition. This is the concrete implementation of our framework. To validate the efficacy of our framework, we evaluate our framework along multiple axes. Behaviorally, the aligned MLLMs improve consistency with human similarity judgements compared to baselines, and retain competitive performance on standard multimodal benchmarks. Beyond behavioral aspect, we conduct searchlight Representational Similarity Analysis (RSA) [Kriegeskorte et al., 2008] using the independent fMRI Natural Scenes Dataset (NSD) [Allen et al., 2022]. The RSA results suggest that the internal representations of our aligned MLLMs demonstrate improved representational similarity to human neural activity. These neural results are not to reveal biological localization, but as a convergent, independent measure of similarity. Our findings demonstrate that explicitly incorporating behavior-derived cognitive priors is an effective and viable pathway for advancing the representational alignment between MLLMs and human cognitive structure.

Our contributions are summarized as follows: (1) We propose a novel alignment framework that moves beyond shallow behavioral mimicry by explicitly internalizing the latent structure of human cognition. By converting behavior-derived semantic dimensions into dimension-grounded linguistic rationales, MLLMs learn human-like decision-making cognition. (2) We introduce a pathway to integrate multidimensional mental representations from cognitive science into MLLMs. (3) Guided by this framework, we construct the alignment dataset that augments human similarity judgements with explicit rationales.

2 Related Work

Recent advances have made significant strides in aligning models with human cognitive behavior. Muttenthaler et al. improve neural network representations using human similarity judgements [Muttenthaler et al., 2023a] and find that model scale and architecture have essentially no effect on the alignment with human behavioral responses, whereas the training dataset and objective function both have a much larger impact [Muttenthaler et al., 2023b]. AligNet [Muttenthaler et al., 2025] distills human similarity judgements from the THINGS dataset [Hebart et al., 2023] into vision models, successfully embedding human-like conceptual hierarchies into visual representations. In the video domain, Garcia and Isik [Garcia and Isik, 2026] introduce behavioral geometric supervision to align video representations with human judgements of social similarity, improving their correspondence with human social perception. Similarly, in the domain of Large Language Models (LLMs), Centaur [Binz et al., 2025] fine-tunes LLMs on an unprecedented scale of human behavioral data (Psych-101), demonstrating remarkable capabilities in predicting and simulating human behavior in any experiment expressible in natural language.

While these pioneering works establish the viability of using human behavioral data for model alignment, they primarily operate under the paradigm of outcome-level behavioral mimicry. By optimizing models directly toward human decision outcomes, these approaches implicitly assume that matching the output distribution entails capturing the underlying cognitive process. However, this leaves a critical gap: it remains opaque whether the models genuinely internalize the latent cognitive structure driving these decisions.

Our work builds on the foundation laid by these efforts but takes a step forward. Rather than relying on outcome matching, we argue that cognitive alignment in MLLMs requires explicit supervision of the reasoning process. By leveraging established behavior-derived semantic dimensions to construct dimension-grounded linguistic rationales, our framework explicitly models why a choice is made. This transforms the implicit cognitive structure into learnable CoT constraints, advancing the alignment paradigm from shallow behavioral mimicry to the integration of latent human cognitive structure.

3 Method

3.1 Triplet-Level Semantic Dimension Attribution

As illustrated in Figure 1, we first build on the established behavior-derived semantic embedding space introduced by Hebart et al. [Hebart et al., 2020, 2023], specifically adopting the object-level semantic dimensions estimated via the Sparse Positive Similarity Embedding (SPoSE) [Zheng et al., 2019] model. To bridge these object representations with each triplet decision, we adopt a Jackknife-based attribution method to identify the most influential dimension underlying human similarity judgement.

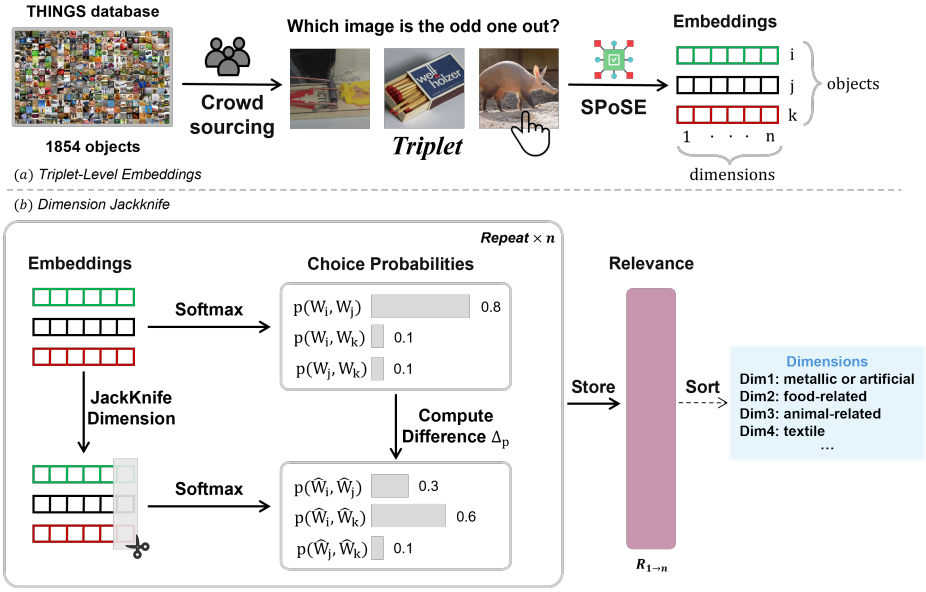


Figure 1: Overview of triplet-level semantic dimension attribution. (a) Human similarity judgements are collected via crowdsourcing. Then learning a representational embedding space via SPoSE yields the multidimensional mental representations of natural objects. (b) For one triplet, we compute the original softmax probability based on the entire representational embedding for all pairs. We then iteratively prune each dimension from the representational embedding space and compute the changes in softmax probabilities relative to those obtained with the full embedding as the relevance score for that dimension. By sorting these relevance scores in descending order, we establish a hierarchical ranking of dimensional importance. The **top-1** dimension yielding the maximum probability change is regarded as the most salient semantic dimension driving the human decision-making.

Object embedding space. As stated in the preliminaries A, given triplet objects $\mathcal{T} = \{i, j, k\}$ (k is odd-one-out), we define the probability that any two objects $\{u, v\} \subset \mathcal{T}$ are identified as the most

similar pair as $p(W_u, W_v)$. This probability is computed via a softmax over the inner products of their embeddings W_u and W_v.

Triplet-level dimensions jackknife. To infer which specific semantic dimension d^* primarily drives a human decision (top-1 dimension), we employ a dimension Jackknife method. We iteratively ablate individual dimensions from the triplet objects embeddings. Let $\hat{W}$ be the embedding vector where the d dimension is pruned. We then recompute the ablated choice probability, denoted as $p(\hat{W}_u, \hat{W}_v)$. The relevance score of dimension d to the human judgement is quantified by the absolute probability difference:

$$\Delta_p = \left| p(W_i, W_j) - p(\hat{W}_i, \hat{W}_j) \right| \tag{1}$$

A larger Δ_p indicates that the original human choice is more sensitive to that pruned dimension. This process is repeated for all $n = 66$ dimensions to yield a relevance scores vector $R_{1\ldots66}$.

3.2 Dimension-Guided Reasoning Enhancement

While the Jackknife attribution effectively isolates the most influential semantic dimension d^* driving human decision-making, utilizing discrete semantic labels (e.g., metallic or artificial) provides a sparse and contextually limited supervision signal. To address this, we convert the discrete semantic dimension d^* label into an explicit, dimension-guided linguistic rationale $r(d^*)$ leveraging the generation capabilities of LLM.

As shown in Figure 2, for one triplet $\mathcal{T}$ and corresponding textual descriptions $cap(\mathcal{T})$, we synthesize the structured prompt 13, $Prompt(\mathcal{T}, cap(\mathcal{T}), d^*)$. Crucially, the LLM is not tasked with predicting the human choice. Instead, it is utilized exclusively as a reasoning enhancement engine. By injecting the semantic dimension d^* into the prompt context, we guide the LLM to articulate exactly how this dimension drives human thinking and decision-making. The LLM processes the prompt to synthesize a coherent CoT reasoning path $r(d^*)$. Ultimately, these dimension-guided rationales provide high-quality, cognitively aligned supervisory data for the MLLMs alignment.

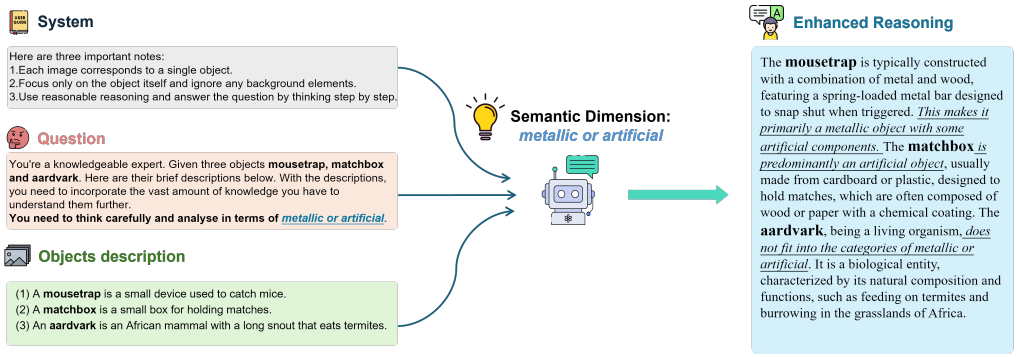


Figure 2: Pipeline for dimension-grounded reasoning enhancement. This pipeline illustrates the process of reasoning enhancement via LLM. For one triplet with their textual descriptions and the most salient semantic dimension d^* driving human decision-making, we integrate them into the structured prompt $Prompt(\mathcal{T}, cap(\mathcal{T}), d^*)$. LLM is then utilized exclusively to enhance reasoning. By injecting the dimension d^*, we guide the LLM to perform targeted reasoning enhancement. This effectively converts discrete cognitive criteria for decision-making into explicit CoT rationale $r(d^*)$.

3.3 Cognitive Latent Structure Alignment

The final stage of our framework is to align MLLMs via Supervised Fine-Tuning (SFT) using the derived explicit linguistic rationale $r(d^*)$. Current alignment paradigms predominantly rely on outcome-level supervision, optimizing models for behavioral mimicry without internalizing the underlying latent structure of human cognition. By explicitly anchoring the MLLMs in human-guided reasoning trajectories, we steer their internal decision-making mechanisms to align with the latent structure of human cognition.

Alignment optimization objective. Given the visual input of triplet objects v and the text instruction x_{inst}, the enhanced reasoning trajectory is denoted as $\tau = [r(d^*); y]$. This formulation naturally factorizes the generative probability into two sequential cognitive phases: rationale generation (thinking) and final decision-making. Let θ parameterize the MLLMs; the alignment objective over our dataset $\mathcal{D}_{\text{O1O}}$ is formulated as:

$$\mathcal{L}_{\text{O1O}}(\theta) = -\mathbb{E}_{(\mathbf{v}, x_{\text{inst}}, \tau) \in \mathcal{D}_{\text{O1O}}} \left[\sum_{t=1}^{\tau_r} \log P_\theta(r_t \mid \mathbf{v}, x_{\text{inst}}, r_{<t}) + \sum_{j=\tau_r+1}^{\tau_r+\tau_y} \log P_\theta(y_j \mid \mathbf{v}, x_{\text{inst}}, \tau_r, y_{<j}) \right] \quad (2)$$

where $P_\theta(\cdot)$ denotes the autoregressive next-token prediction probability parameterized by the MLLMs. In this formulation, τ_r and τ_y denote the sequence lengths of the rationale and the final decision-making, respectively. During the initial reasoning phase ($t \in [1, \tau_r]$), r_t represents the t-th rationale token generated conditioned on the visual inputs $\mathbf{v}$, the text instruction x_{inst}, and the preceding tokens $r_{<t}$. In the subsequent decision-making phase ($j \in [\tau_r + 1, \tau_r + \tau_y]$), y_j denotes the j-th target token. Crucially, this second phase is conditioned on τ_r, which herein represents the fully generated rationale sequence acting as a cognitive anchor. By optimizing this negative log-likelihood, we constrain the MLLMs internal decision-making mechanisms to internalize the latent human-like structure before the final decision.

Joint optimization. While $\mathcal{L}_{\text{O1O}}(\theta)$ effectively instills cognitive latent structure, fine-tuning exclusively on a narrow task distribution risks Catastrophic Forgetting (CF). To preserve foundational capabilities while aligning, we employ a joint optimization strategy. We use multi-task objective by mixing $\mathcal{D}_{\text{O1O}}$ with a general-domain dataset $\mathcal{D}_{\text{gen}}$ (e.g., visual question answering data):

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{O1O}}(\theta) + \lambda \mathcal{L}_{\text{gen}}(\theta) \quad (3)$$

where $\mathcal{L}_{\text{gen}}(\theta)$ denotes the standard autoregressive next-token prediction loss on general multimodal data, ensuring the model retains its foundational generative capabilities. λ is a regularization hyperparameter, implemented via dynamic data sampling ratios during training (see Appendix G.4 for details). This multi-objective formulation allows the MLLMs to internalize human-consistent reasoning while retaining their broad generalization.

3.4 Neural Searchlight RSA

To provide a convergent, independent measure of representational alignment beyond behavioral evaluations, we conduct searchlight RSA using NSD. This neural evaluation is not intended to claim precise biological localization of MLLMs features, but rather to serve as a robust, modality-independent validation of our alignment framework. Given the same visual stimuli $\mathcal{S}$ for both MLLMs and human, we denote this as $\mathcal{S} = \{s_1, s_2 \ldots, s_n\}$ and compute Representational Dissimilarity Matrix (RDM).

MLLMs RDM. After extracting the internal representations via a forward propagation (detailed in Appendix F.1), we denote $h_i \in \mathbb{R}^D$ as the D-dimensional hidden state vector extracted from the <|vision_end|> token for s_i. We quantify the internal representational geometry by computing the MLLMs RDM, denoted as $RDM^M \in \mathbb{R}^{S \times S}$. The pairwise representational dissimilarity between any two stimuli s_i and s_j is defined as:

$$RDM^M_{i,j} = 1 - \rho(h_i, h_j) \quad (4)$$

where $\rho(\cdot, \cdot)$ is the Pearson correlation coefficient across the D feature dimensions.

Human RDM. For a spherical searchlight neighborhood centered at voxel c comprising C contiguous voxels (radius = 3 voxels), let $b^c_i \in \mathbb{R}^C$ denote the local multi-voxel neural activation pattern elicited by stimulus s_i. The corresponding localized RDM, denoted as $RDM^c \in \mathbb{R}^{S \times S}$, is computed by:

$$RDM^c_{i,j} = 1 - \rho(b^c_i, b^c_j) \quad (5)$$

Searchlight RSA score. To quantify the representational alignment between the MLLMs internal representation space and the localized human cortex, we compute the voxel-wise neural alignment score, $RSA(\mathcal{M}, c)$. Let $u(\cdot)$ denote the operator that extracts the flattened upper triangular elements of a given RDM, excluding the main diagonal. The alignment score at searchlight center c is defined as the Spearman correlation coefficient between the respective upper triangular vectors:

$$RSA(\mathcal{M}, c) = \rho_s(u(RDM^M), u(RDM^c)) \tag{6}$$

To evaluate the representational enhancement driven by our alignment framework, we compute the voxel-wise alignment shift, denoted as $\Delta_{RSA}(c)$. This shift quantifies the difference in alignment scores between the aligned MLLMs ($\mathcal{M}_{align}$) and the unaligned baseline ($\mathcal{M}_{base}$):

$$\Delta_{RSA}(c) = RSA(\mathcal{M}_{align}, c) - RSA(\mathcal{M}_{base}, c) \tag{7}$$

Prior to cortical projection, we apply a lower-bound threshold to the alignment shift map to filter out negligible negative fluctuations. Empirically, negative $\Delta_{RSA}(c)$ values are extremely close to zero and arise from computational noise rather than meaningful representational degradation. Therefore, we clamp these values to a near-zero floor ($\epsilon = 10^{-8}$):

$$\tilde{\Delta}_{RSA}(c) = \max(\epsilon, \Delta_{RSA}(c)) \tag{8}$$

The resulting non-negative alignment shift map, $\tilde{\Delta}_{RSA}$, is then projected onto a flattened cortical surface using pycortex [Gao et al., 2015].

4 Experiments

4.1 Dataset

We construct a hybrid dataset including $\mathcal{D}_{\text{OIO}}$ and $\mathcal{D}_{\text{gen}}$. The composition is summarized in Figure 3 (a) and examples are shown in Figure 3 (b). Details are provided in Appendix G.

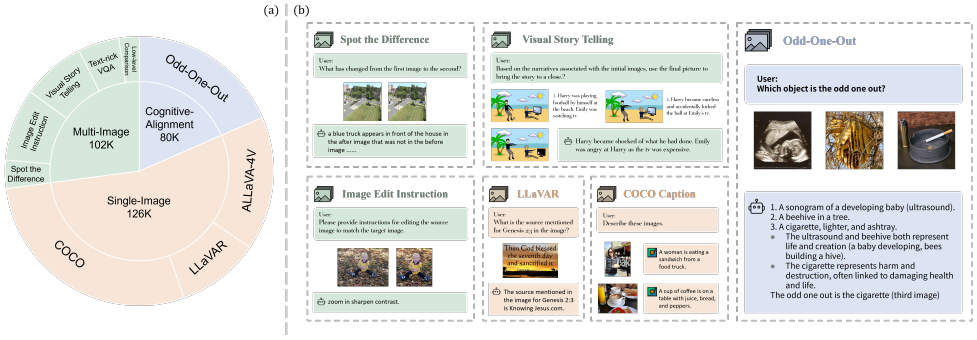


Figure 3: (a) Overview of hybrid dataset. The total dataset includes three subsets: Single-image (126K samples), Multi-image (102K samples), and Cognitive-Alignment (80K samples). (b) Examples. The examples include Spot the Difference, Image Edit Instruction, Visual Storytelling, LLaVAR, COCO Caption, and Odd-One-Out tasks.

4.2 Setup

We conduct experiments across the Qwen3-VL [Bai et al., 2025a] model family, specifically utilizing the ∼4B and ∼8B instruction models, as well as the ∼30B Mixture-of-Experts (MoE) model. We denote the aligned models fine-tuned via our framework as **CogAlign**, and their unaligned counterparts as **Baseline**. Full implementation details are provided in Appendix H.1.

In addition to these MLLMs, we incorporate **Centaur**, a foundation model of human cognition. This serves as a specialized external reference, allowing us to benchmark our aligned models against an architecture optimized for cognitive tasks.

4.3 Main Result

Generalization performance. A critical concern in model alignment is the potential degradation of general capabilities. To investigate whether aligning with the latent structures of human cognition compromises broader reasoning abilities, we evaluate both the baseline and aligned models on two multimodal benchmarks: MMMU [Yue et al., 2024] (val) and MMBench [Liu et al., 2024] (en, test).

As reported in Table 1, we evaluate the general capabilities to investigate the potential alignment tax. The results indicate a trade-off: while the aligned models achieve consistent performance gains across all scales on MMMU_{val} benchmark (e.g., +4.34 improvement for Qwen3-VL-30B), they exhibit a minor degradation on $\text{MMBench}_{\text{en, test}}$ compared to their unaligned baselines. This suggests that aligning with human cognitive latent structure incurs a slight alignment tax on general-purpose multimodal tasks.

Table 1: Performance comparison of baseline and aligned models. CogAlign yields consistent accuracy improvements on MMMU_{val}, while incurring a slight performance degradation on $\text{MMBench}_{\text{en, test}}$ across all model scales. Details are provided in Table 9 (Appendix H.3).

Model	MMMU_{val} (%)		$\text{MMBench}_{en,test}$ (%)	
Qwen3-VL	Baseline	CogAlign	Baseline	CogAlign
4B	48.44	**51.56**	**82.96**	81.39
8B	50.67	**54.22**	**83.91**	83.86
30B-A3B	53.22	**57.56**	**87.44**	83.86

Human consistency. We quantify human consistency using accuracy on the O1O validation set comprising 453,642 held-out triplets, where higher accuracy indicates better alignment. This massive scale ensures that our evaluation captures robust results rather than anecdotal success.

As detailed in Table 2, we evaluate the models against two distinct baseline settings to establish a comprehensive comparison. In the Direct prompt 15 setting, the model is prompted to output the final option (e.g., A, B, or C) immediately. In the CoT prompt 16 setting, the model is required to think step-by-step before decision-making. The results demonstrate that our alignment framework yields improvements over both baseline settings and the performance of aligned models remains stable across all scales (ranging from 56.21 to 56.99). While the aligned models outperform not only the Direct baselines but also more capable CoT baselines across all models, the relative gain over the CoT baseline narrows as model capacity increases. This shrinking delta is due to the upward trajectory of the baseline itself.

Table 2: Human consistency evaluation across baseline and aligned models. The baseline models are evaluated under two inference paradigms: Direct (predicting option immediately) and CoT (thinking step-by-step prior to decision-making). The proposed alignment framework consistently outperforms both the Direct and the more capable CoT baselines across all model scales (noise ceiling is 68.74 [Hebart et al., 2023]).

Model		Baseline (%)		CogAlign (%)
Centaur	70B	51.96		—
		Direct	CoT	Ours
Qwen3-VL	4B	41.48 ± 0.30	53.37 ± 0.30	**56.81** ± 0.29
	8B	40.46 ± 0.29	52.28 ± 0.30	**56.99** ± 0.29
	30B-A3B	37.09 ± 0.28	55.38 ± 0.30	**56.21** ± 0.29
Noise Ceiling		68.74 ± 1.07		

To rigorously validate these gains, we report 95% Confidence Intervals (CIs) estimated via bootstrap resampling. For the Qwen3-VL series, the CIs of the aligned models (e.g., $[56.70, 57.28]$ for the $\sim$8B model) do not overlap with those of their respective baselines, indicating a statistically significant enhancement in human-model consistency by a McNemar test [McNemar, 1947] ($p < 0.001$).

Neuroscientific analysis. To investigate whether our framework improves the alignment between the internal representational geometry of MLLMs and the human brain, we analyze the neural align-

ment shift ($\tilde{\Delta}_{RSA}(c)$). Following the method detailed in Section 3.4, this shift is computed using the concatenated hidden states from the MLLMs last four layers and visualized via cortical projection.

As illustrated in Figures 4 and 5, the resulting cortical heatmaps reveal consistent representational alignment improvements across all subjects (subject 1, 2, 5, and 7 [Takagi and Nishimoto, 2023, Scotti et al., 2023]; paired t-test, $p < 0.001$). Detailed discussion of these results is provided in the Appendix H.6.

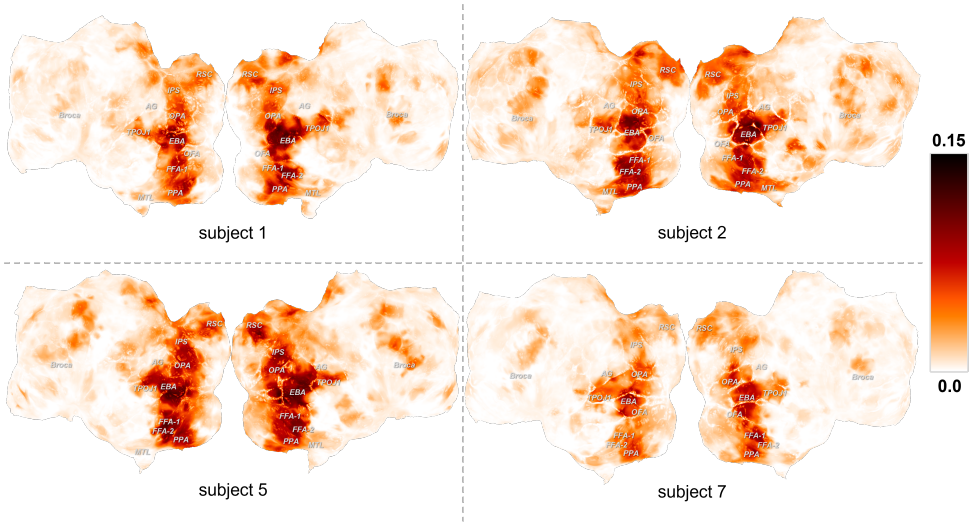


Figure 4: Cortical projection of representational alignment improvements across subjects. The heatmaps visualize the voxel-wise difference in $\tilde{\Delta}_{RSA}(c)$ scores between the aligned and baseline Qwen3-VL-4B-Instruct model. Darker colors indicate better representational alignment between models and human brain. Consistent positive shifts are observed, demonstrating that our alignment framework robustly shifts the MLLMs internal representational geometry to be more human-like.

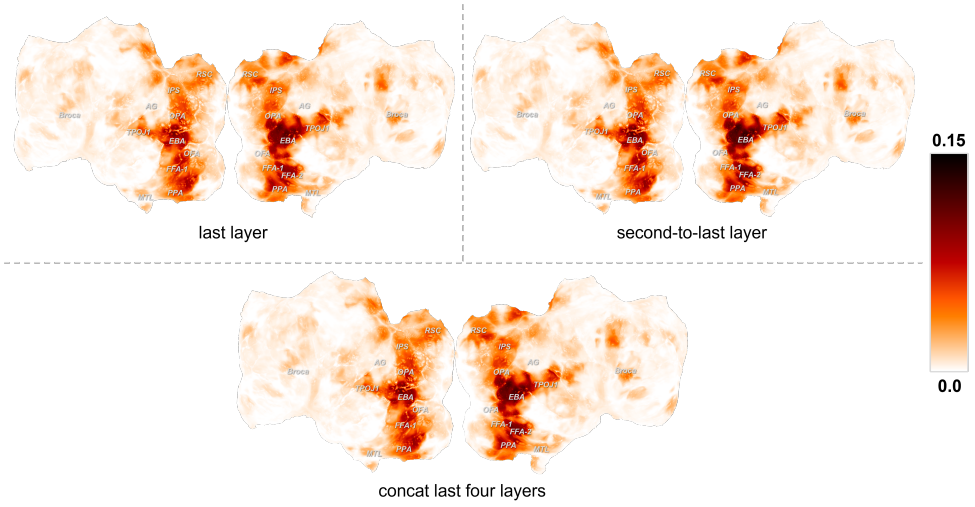


Figure 5: Stability of alignment shifts across different extraction depths. The cortical projections display the alignment improvements $\tilde{\Delta}_{RSA}(c)$ of subject 1 computed using features from the last layer, the second-to-last layer, and the concatenation of the last four layers. Darker regions reflect better alignment improvement. The spatial distribution of these positive shifts remains consistent regardless of the extraction depth, confirming that the achieved representational alignment with human cognition is robust and not a selective layer choice.

4.4 Ablation Result

Necessity of alignment over prompting. To determine whether MLLMs can show human-like cognitive structures purely via inference-time prompting or if our alignment is necessary, we ablate the Qwen3-VL-4B-Instruct baseline against three prompting strategies: (1) Standard CoT (S-CoT 16), which uses unconstrained step-by-step reasoning. (2) d^*-Guided CoT (d^*-CoT 17), which strictly constrains reasoning to the most salient semantic dimension (d^*). (3) Full-Space CoT (FS-CoT 18), which requires dynamic grounding one or more dimension(s) across the complete 66-dimensional behavioral semantic space.

As shown in Table 3, we compare inference-time prompting with our alignment framework on Qwen3-VL Instruct model. Although explicitly providing semantic dimensions enables the model to access these concepts during reasoning, it does not guarantee stable reasoning over these dimensions. Neither d^*-CoT nor FS-CoT consistently improves over standard reasoning (S-CoT). d^*-CoT reduces performance from 53.37 to 49.01, suggesting that imposing a cognitive dimension may constrain the model's decision process without ensuring effective utilization of the underlying semantic information. Similarly, FS-CoT achieves 52.70, remaining comparable to S-CoT. In contrast, our CogAlign achieves 56.81, exceeding S-CoT and FS-CoT by 3.44 and 4.11 percentage points, respectively. Our framework incorporates these latent cognitive structures into the model, enabling more consistent human-like decision patterns.

Table 3: Prompting versus alignment. We compare three inference-time prompting strategies (S-CoT, d^*-CoT, and FS-CoT) with our alignment framework. Notably, providing semantic dimensions in the prompt during inference time negatively impacts model performance, whereas our CogAlign achieves the best overall results. The chance level is 33.33%. See Appendix H.4 for details.

Model		S-CoT	d^*-CoT	FS-CoT	CogAlign
Qwen3-VL	4B	53.37	49.01	52.70	**56.81**

Generalizability across MLLMs families. To verify that our alignment framework is architecture-agnostic, we extend our evaluation to Qwen2.5-VL [Bai et al., 2025b] and Gemma3 [Kamath et al., 2025] instruction models. To balance computational efficiency with reliability, we conduct this ablation on a subset (∼75.6K triplets, 1/6) of the O1O validation set.

As reported in Table 4, our framework yields consistent improvements in human behavioral consistency across distinct model families. The aligned Qwen2.5-VL and Gemma3 outperform their respective CoT baselines by 2.66 and 1.51.

Table 4: Ablation on architectural generalizability.

Model		Baseline (CoT)	CogAlign (Ours)
Qwen2.5-VL	7B	56.63	**59.29**
Gemma3	12B	55.77	**57.28**

Similarity of behavioral representational geometry. Beyond agreement on triplet choices, we examine whether alignment captures the broader organization of human similarity judgements.

Choice-based representational similarity matrices (RSMs) represent pairwise similarity relationships among objects, allowing us to assess model-human correspondence at the level of behavioral representational geometry. We compute the Pearson correlation of the choice-based RSMs derived from Qwen2.5-VL-7B and human on a 48-object subset (∼91.56K triplets) drawn from the complete dataset. As shown in Figure 6, the aligned model achieves 80.93% exceeding its CoT baseline 73.30% and Centaur 65.13%. These results suggest that alignment improves correspondence with human behavioral representational geometry beyond triplet choices. Details are provided in Appendix H.7.

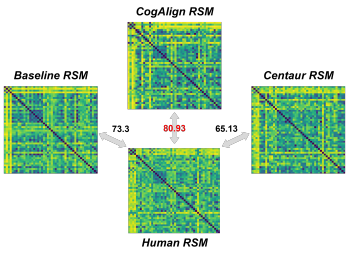


Figure 6: Comparison of choice-based RSMs between Qwen2.5-VL-7B and human.

5 Conclusion

In this paper, we present a data-driven framework to align MLLMs with the latent structure of human cognition via behavior-derived semantic dimensions. Moving beyond outcome-level mimicry,

our framework translates the multidimensional mental representations of natural objects underlying human similarity judgements into explicit linguistic rationales, enabling MLLMs to internalize these priors. Experimental evaluations demonstrate the efficacy of our alignment framework. The aligned MLLMs achieve improvements in behavioral consistency with human. Ablation studies reveal that these cognitive priors should be internalized: providing semantic dimensions via inference-time prompting paradoxically degrades performance, suggesting a fundamental gap between prompting and our framework. Furthermore, independent searchlight RSA indicates that the MLLMs internal representations become more aligned with human neural activity after aligning. While a minor trade-off is observed on general benchmarks, this work provides a proof-of-concept for bridging cognitive science and MLLMs alignment. Our framework points to a data-driven route for incorporating human cognitive structure into MLLMs alignment. We hope this work inspires further exploration toward developing more controllable, transparent, and human-like artificial cognitive systems. Furthermore, to clarify the commensurability of our baselines and explicitly outline the value of this process-level alignment framework for both the machine learning and cognitive science communities, we provide extended discussion in Appendix E.

6 Limitations

While our framework improves alignment between MLLMs and the latent structure of human cognition, we note some limitations.

Our triplet-level semantic dimension attribution relies on a top-1 approximation, identifying a single dominant dimension (d^*). This isolates the primary semantic axis, but does not model the joint contributions of multiple dimensions or their context-dependent interactions. The resulting supervision may therefore simplify the structure underlying human similarity judgements.

The 66 behavior-derived semantic dimensions are object-centric, characterizing perceptual and conceptual properties primarily. Although this space includes some abstract dimensions, such as valuable or precious and weapon-related (Table 12), their presence does not establish coverage of the abstractions required for mathematical reasoning or reasoning about complex geometric and topological structures. Consequently, the applicability of our alignment framework to highly abstract multimodal domains remains to be further explored (see Appendix B for details).

As with many synthesized datasets, the rationales generated in our pipeline are not flawless. While our framework strictly grounds every rationale in a valid, human-derived cognitive dimension, the subsequent verbalization by the LLMs can occasionally introduce factual inconsistencies or ad hoc associations. We provide transparent and detailed discussion in Appendix C.

The improvement over the CoT baseline becomes smaller for larger models. This trend is driven by the emergent conceptual structures inherent in stronger models, which raises the initial untuned performance closer to our current alignment ceiling (see Appendix D for details).

Acknowledgments and Disclosure of Funding

This work was supported in part by Beijing Natural Science Foundation under Grant L243016; in part by the National Natural Science Foundation of China under Grant 62576336 and 82272072; and in part by Zhongguancun Academy (Grant No.s P210260301).

References

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

GLM-5-Team: Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Changde Du, Kaicheng Fu, Bincheng Wen, Yi Sun, Jie Peng, Wei Wei, Ying Gao, Shengpei Wang, Chuncheng Zhang, Jinpeng Li, et al. Human-like object concept representations emerge naturally in multimodal large language models. *Nature Machine Intelligence*, 7(6):860–875, June 2025.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, pages 50742–50768, 2023.

Shobhita Sundaram, Stephanie Fu, Lukas Muttenthaler, Netanel Tamir, Lucy Chai, Simon Kornblith, Trevor Darrell, and Phillip Isola. When does perceptual alignment benefit vision representations? In *Advances in Neural Information Processing Systems*, volume 37, pages 55314–55341, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL `https://arxiv.org/abs/1707.06347`.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741, 2023.

Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, and Chris I. Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4 (11):1173–1185, November 2020.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, February 2023.

Florian P. Mahner, Lukas Muttenthaler, Umut Güçlü, and Martin N. Hebart. Dimensions underlying the representational alignment of deep neural networks with humans. *Nature Machine Intelligence*, 7(6):848–859, June 2025.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 2008.

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022.

Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, Katherine Hermann, Andrew Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. In *Advances in Neural Information Processing Systems*, volume 36, pages 50978–51007, 2023a.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023b.

Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C. Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K. Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647(8089):349–355, November 2025.

Kathy Garcia and Leyla Isik. Behavioral geometric supervision aligns video foundation models with human social perception, 2026. URL `https://arxiv.org/abs/2510.01502`.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, August 2025.

Charles Y. Zheng, Francisco Pereira, Chris I. Baker, and Martin N. Hebart. Revealing interpretable object representations from human behavior. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

James S. Gao, Alexander G. Huth, Mark D. Lescroart, and Jack L. Gallant. Pycortex: an interactive surface visualizer for fmri. *Frontiers in Neuroinformatics*, 9, 2015.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, June 2024.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064, pages 216–233, 2024.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, June 1947.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14453–14463, 2023.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, ethan cohen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. In *Advances in Neural Information Processing Systems*, volume 36, pages 24705–24728, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report, 2025b. URL `https://arxiv.org/abs/2502.13923`.

Gemma Team: Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, et al. Gemma 3 technical report, 2025. URL `https://arxiv.org/abs/2503.19786`.

Alfonso Caramazza and Jennifer R. Shelton. Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience*, 10(1):1–34, 1998.

Linda L. Chao, James V. Haxby, and Alex Martin. Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience*, 2(10):913–919, October 1999.

Talia Konkle and Aude Oliva. Canonical visual size for real-world objects. *Journal of experimental psychology. Human perception and performance*, 37(1):23–37, 2011.

Brian Murphy, Partha Talukdar, and Tom Mitchell. Learning effective and interpretable semantic models using non-negative sparse embedding. In *Proceedings of COLING 2012*, pages 1933–1950, December 2012.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37:547–559, 2005.

Kimi Team: Tongtong Bai, Yifan Bai, Yiping Bao, M. C., Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y. Charles, et al. Kimi k3: Open frontier intelligence, 2026. URL `https://arxiv.org/abs/2607.24653`.

Qwen: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL `https://arxiv.org/abs/2606.19348`.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025a. URL `https://arxiv.org/abs/2506.05176`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, June 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, 2014.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model, 2024. URL https://arxiv.org/abs/2402.11684.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding, 2024. URL https://arxiv.org/abs/2306.17107.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun MA, and Chunyuan Li. Llava-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pages 81182–81199, 2025.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4024–4034, October-November 2018.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*, volume 36, pages 31428–31449, 2023.

Rumeysa Bodur, Erhan Gundogdu, Binod Bhattarai, Tae-Kyun Kim, Michael Donoser, and Loris Bazzani. iedit: Localised text-guided image editing with weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7426–7435, June 2024.

Hareesh Ravi, Kushal Kafle, Scott Cohen, Jonathan Brandt, and Mubbasir Kapadia. Aesop: Abstract encoding of stories, objects, and pictures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2052–2063, October 2021.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, June 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformers Reinforcement Learning, 2020. URL https://github.com/huggingface/trl.

QwenLM. Qwen3-vl, 2026. URL https://github.com/QwenLM/Qwen3-VL/tree/main.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models, 2020. URL https://arxiv.org/abs/1910.02054.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

ArdalanM. Understanding the gap in benchmark results for qwen3vl models, 2025. URL https://github.com/EvolvingLMMs-Lab/lmms-eval/issues/901.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. LMMs-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, April 2025b.

Bo Li*, Peiyuan Zhang*, Kaichen Zhang*, Fanyi Pu*, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, et al. Lmms-eval: Accelerating the development of large multimodal models, March 2024. URL https://github.com/EvolvingLMMs-Lab/lmms-eval.

A Preliminaries

Prior experimental, neuropsychological and neuroimaging evidence have led to the proposal of object dimensions such as animacy, manipulability and real-world size [Caramazza and Shelton, 1998, Chao et al., 1999, Konkle and Oliva, 2011], but these dimensions describe only a selective and largely incomplete portion of our mental representational space. Our work builds on a previously established behavior-derived semantic embedding space, which provides visual and conceptual mental representations for a wide, representative range of natural objects [Hebart et al., 2020, 2023]. Specifically, this space maps a set of 1854 object concepts into a sparse and non-negative vector space, where each dimension corresponds to a distinct conceptual and perceptual property (e.g., metallic, animal-related, textile).

A.1 Sparse Positive Similarity Embedding (SPoSE)

We introduce the SPoSE proposed by Charles Y. Zheng et al. [Zheng et al., 2019] in this section. SPoSE is based on three assumptions.

Semantic representation of concepts. The first assumption is that each object concept c_i can be represented by semantic features, quantified in a vector $\mathbf{x_i} = (x_{i1}, \ldots, x_{ip})$. These features correspond dimensions of a latent space where the geometrical relationships encode conceptual (dis)similarity. The SPoSE model estimates these features *from behavioral data alone* using the probabilistic model. To ensure interpretability, the SPoSE model constrains each dimension x_{if} in the vector $\mathbf{x_i}$ to be real and non-negative, reflecting the degree to which the aspect of meaning it represents is present and influences subject behavior [Murphy et al., 2012]. Furthermore, a sparsity constraint is added to the SPoSE model for sparse these dimensions [McRae et al., 2005].

Probabilistic model. The second assumption is that the decision in a triplet task trial is explained by the similarity between the embedding vectors of the presented concepts. For any two concepts c_i and c_j, the SPoSE model defines the similarity S_{ij} as the dot product of their corresponding vectors:

$$S_{ij} = \langle \mathbf{x_i}, \mathbf{x_j} \rangle = \sum_{f=1}^{p} x_{if} x_{jf}. \tag{9}$$

The third assumption models the outcome of the triplet task as a stochastic process. This assumption reflects both trial-to-trial inconsistencies in how an individual subject makes the decision, as well as subject differences in the decision-making process. To account for behavioral inconsistencies, the probability of a subject choosing a specific pair is modeled by Softmax Action Selection:

$$\Pr[y_1, y_2] \propto \exp(S_{y_1,y_2}) \tag{10}$$

Since probabilities of the three possible outcomes add to one, this gives

$$\Pr[y_1, y_2] = \frac{\exp(S_{y_1,y_2})}{\exp(S_{i_1,i_2}) + \exp(S_{i_1,i_3}) + \exp(S_{i_2,i_3})} \tag{11}$$

In this formulation, the vectors $\mathbf{x_i}$ are unnormalized. This design choice allows vector similarities to be arbitrarily large, modeling situations where decision is near-deterministic (e.g., when two concepts are virtually identical).

Parameter fitting. The embedding vectors $(\mathbf{x_1}, \ldots, \mathbf{x_m})$ are inferred by minimizing a regularized maximum likelihood objective function. For the given triplets datasets (total number is n), where $(i_{1,j}, i_{2,j}, i_{3,j})$ denotes the concepts in the j-th triplet and $(y_{1,j}, y_{2,j})$ indicates the chosen pair. The objective function to be minimized is defined as:

$$\sum_{j=1}^{n_{train}} \log \left(\frac{\exp({\mathbf{x_{y_{1,j}}}}^T \mathbf{x_{y_{2,j}}})}{\exp({\mathbf{x_{i_{1,j}}}}^T \mathbf{x_{i_{2,j}}}) + \exp({\mathbf{x_{i_{1,j}}}}^T \mathbf{x_{i_{3,j}}}) + \exp({\mathbf{x_{i_{2,j}}}}^T \mathbf{x_{i_{3,j}}})} \right) + \lambda \sum_{i=1}^{m} ||\mathbf{x_i}||_1 \tag{12}$$

where p is specified number of dimensions, n_{train} is the number of training triplets. $|| \cdot ||_1$ is the L1 norm, so $||\mathbf{x}||_1 = \sum_{f=1}^{p} |x_f|$, and $x_f \geq 0$ for $f = 1, \ldots, p$. λ is the regularization parameter.

A.2 Behavior-Derived Embedding Space via SPoSE

The object representations are learned from human similarity judgement behavioral data using the SPoSE model. In the triplet O1O task, participants are presented with three objects, i, j, and k, and asked to identify the most dissimilar one. Assuming k is chosen as the odd-one-out, the psychological similarity between i and j is considered the highest among the pairs. SPoSE models this by assigning a non-negative latent vector $\mathbf{x} \in \mathbb{R}^d_{\geq 0}$ to each object. The probability of choosing k as the odd-one-out is formulated using a softmax over the inner products of the object vectors:

$$P(k \mid i, j, k) = \frac{\exp(\mathbf{x}_i^\top \mathbf{x}_j)}{\exp(\mathbf{x}_i^\top \mathbf{x}_j) + \exp(\mathbf{x}_i^\top \mathbf{x}_k) + \exp(\mathbf{x}_j^\top \mathbf{x}_k)} \tag{13}$$

By applying L_1 regularization during optimization, SPoSE enforces sparsity and non-negativity, ensuring that the resulting dimensions are interpretable as the latent structure of human cognition [Hebart et al., 2020].

A.3 Justification for 66 Dimensions

The critical property of this embedding space is its dimensionality, **d = 66**. In many representation learning contexts, dimensionality is treated as an arbitrary hyperparameter tuned for downstream performance. However, in this behavior-derived space, $d = 66$ is saturation. Analysis conducted by Hebart et al. [Hebart et al., 2023] demonstrates that while the number of reproducible dimensions scales with the volume of behavioral data, it ultimately saturates around 67.5 dimensions and would reach ~ 66.5 dimensions for $4.5 - 5$ million trials. Indeed, when re-running the SPoSE model with the full dataset of 4.70 million trials (4.10 million for training), embedding dimensionality turns out as predicted: from a set of 72 randomly-initialized models, they choose the most reliable embedding as the final embedding, revealing 66 interpretable dimensions underlying perceived similarity judgements. Thus, increasing dataset size beyond this large dataset may no longer yield noticeable improvements in predictive performance or changes in embedding dimensionality at the global level of similarity, and potential improvements may not justify the cost of collecting additional data. These 66 dimensions represent a data-driven, globally saturated approximation of the latent structure of human cognition used to conceptualize this broad set of natural objects. This provides a rigorous and interpretable foundation for our proposed alignment framework.

B Impact of Incorporating Abstract Dimensions

Our alignment framework makes no assumptions about the semantic content of the dimensions. Incorporating richer abstract dimensions therefore requires replacing or augmenting the embedding space rather than modifying the core framework.

We envision three pathways for this integration: (a) Swapping the behavioral substrate. Applying the identical pipeline to O1O data collected on stimuli where abstract criteria dominate (e.g., social or moral scenarios). (b) Augmenting the existing space. Concatenating normative ratings for objects along abstract attributes with the SPoSE dimensions before Jackknife attribution, allowing the attribution step to decide whether an abstract or object-centric criterion best explains the human choice. (c) Context-conditioned judgements. Since abstract criteria often become decision-relevant only under specific contexts, collecting context-conditioned O1O data and injecting this context into the rationale prompt.

Abstract criteria (e.g., social acceptability) are substantially more subject- and culture-dependent, introducing variability into the decision-making process. While this could enhance the model's alignment with complex human values, it also raises the bias-auditing concerns we raise in the Impact Statement. As illustrated by the emergent stereotypical gender dimensions, behaviorally derived abstract axes can encode societal norms or biases that human may not wish to instill wholesale, necessitating careful curation when expanding the semantic space.

C Rationale Quality Analysis

In our framework, the supervision signal originates from two distinct sources. (1) Human Population: The decisive supervisory anchor, the choice and its operative cognitive dimension, is attributed from

4.70M validated human judgements via SPoSE and Jackknife. (2) LLM Generator: The LLM role is strictly restricted to verbalizing this behaviorally grounded criterion without altering the underlying decision.

A concern regarding this pipeline is the quality of the LLM-generated rationales. Since no generation is flawless, we conduct a transparent diagnostic of the rationale quality, evaluate the robustness of the framework across different rationale generators, and clarify why the framework yields robust gains despite data imperfections.

Independent MLLM-as-judge diagnostic. To quantitatively evaluate rationale quality without self-preference bias, we employ a stronger, independent MLLM (Kimi-K3 [Bai et al., 2026]) as a rubric-based judge on a random sample of training rationales. Given the triplet images, the corresponding dimension, and the rationale, the judge scores four independent criteria:

- **Dimension faithfulness (D):** The given dimension is the actual operative basis of the comparison.
- **Factual correctness (F):** Claims about the objects are visually supported or represent stable common knowledge.
- **Logical validity (L):** The stated premises consistently entail the odd-one-out choice.
- **Non-forcedness (N):** The dimension-object relation is conventional and discriminative, penalizing ad hoc associations.

As shown in Table 5, for Qwen2.5-72B-Instruct generator, the independent MLLM-as-judge diagnostic reveals that while 93.37% of rationales are faithfully grounded in the assigned dimension, only about 70.45% are strictly logically valid, 73.11% are factually flawless, and 54.73% are completely non-forced. Despite these verbalization imperfections, our results confirm that the framework consistently improves alignment across multiple MLLMs and generator LLMs.

Table 5: Rationale quality across generators.

Generator	D	F	L	N
Qwen2.5-72B-Instruct (Ours)	93.37	73.11	70.45	54.73
DeepSeek-V4-Flash	96.21	81.06	69.89	62.12

Cross-generator semantic convergence and robustness. To ensure that the framework effectiveness does not rely on the idiosyncrasies of a single generator (Qwen2.5-72B [Yang et al., 2025]), we regenerate the full set rationales using DeepSeek-V4-Flash [DeepSeek-AI et al., 2026] under identical prompts.

Semantic convergence: Using Qwen3-Embedding-0.6B [Zhang et al., 2025a], we compute the pairwise semantic similarity between rationales produced by the two generators for identical triplets. The high similarity (Mean = 83.39%, Std = 5.93%) confirms that different LLMs converge on similar reasoning pathways when anchored by the same cognitive dimension.

Performance robustness: We fine-tune the Qwen3-VL-4B-Instruct using the generated DeepSeek-V4-Flash rationales. The accuracy (56.52%) closely matches our original result using Qwen2.5-72B rationales (56.81%), and exceeds the CoT baseline (53.37%).

Why imperfect rationales drive consistent improvement. A question arises: if a substantial proportion of training rationales are identified as flawed or forced, why should they be included, and why do they still drive performance improvements?

We argue that this highlights a core strength of our approach. The foundational anchor of every sample remains a valid human consensus. Even when the LLM verbalization is imperfect or slightly forced, it still forces the model to map its internal representations toward a validated human cognitive dimension rather than relying on uninterpretable statistical shortcuts. Crucially, our experiments demonstrate that the performance gains are the property of this dual-supervision framework. The

exact same unfiltered dataset yields consistent improvements across five different MLLMs configurations spanning multiple scales and architectures (e.g., Qwen3-VL, Qwen2.5-VL, Gemma3 in Tables 2 and 4). The fact that alignment gains replicate across multiple architectures and independent rationale generators, despite the quantified imperfections, provides strong evidence that the improvement stems from the structure of the framework itself rather than data curation.

While employing a closed-loop validation procedure or a frontier model (like Kimi-K3) to regenerate and filter the dataset would undoubtedly improve data quality and aligned model performance, such large-scale regeneration is often economically prohibitive for exploratory studies. Our findings confirm that even with accessible open-weight generators and inherently noisy verbalizations, injecting human cognitive priors explicitly via this framework improves alignment.

D Scaling Dynamics and Alignment Ceiling Analysis

As observed in our main results (Table 2), the relative accuracy gain of our aligned model over the CoT baseline diminishes as the parameter size of the MLLM increases. We decompose this scaling trend to provide more detailed discussion.

Across all model scales, the absolute performance of our aligned models remains remarkably stable: 56.81 (4B), 56.99 (8B), and 56.21 (30B). The shrinking delta is driven by the rising CoT baseline. Stronger MLLMs emerge human-like conceptual structures and world knowledge during large-scale pre-training. Consequently, the untuned starting point of larger models naturally moves closer to the human-like behavioral target, leaving less accessible headroom for our current alignment formulation. Currently, our pipeline only relies on top-1 dimension, supervising the model with a single dominant criterion. However, human similarity judgements are often nuanced, involving complex trade-offs among multiple cognitive dimensions. Supervising only the primary axis creates an implicit upper bound on task performance. Furthermore, our current alignment dataset comprises 80K sampled triplets, representing only $\sim$2% of the 4.1M available training triplets.

This yields a concrete, testable prediction for future alignment research: enriching the supervision signal through multidimensional rationales and scaling the alignment data far beyond the current $\sim$2% sample should effectively break the ceiling. As model capabilities increase, they possess the representational capacity to absorb these denser, multi-faceted cognitive priors.

E Framework Scope and Interdisciplinary Value

We further clarify the conceptual scope of our alignment framework, delineate its distinction from existing baselines, and outline its specific utility for researchers.

Baseline commensurability and process-level internalization. A core contribution of our work is shifting the alignment pathway for MLLMs from outcome-level behavioral mimicry to the process-level internalization of latent human cognitive structures. This distinction determines which baselines are commensurable with our approach. Post-hoc feature reweighting and selection methods operate as readouts on frozen feature spaces, producing task-specific discriminative mappings. They do not modify the model's internal reasoning and they are difficult to apply to the autoregressive generative setting of MLLMs. Consequently, the appropriate same-class baseline is a behaviorally fine-tuned model.

As reported in our main results, we compare against Centaur, which is fine-tuned on the largest existing human behavioral corpus (Psych-101). Our aligned models outperform Centaur both behaviorally and in representational similarity in Table 2 and Figure 6. Furthermore our prompting ablations in Table 3 function as a non-parametric analogue to feature selection. This demonstrates that cognitive criteria should be internalized through joint rationale-decision supervision, rather than merely provided as external features.

Broad use cases and value for researchers. Beyond empirical performance, our framework offers a general, data-efficient recipe for injecting population-level cognitive priors into MLLMs without requiring architectural modifications.

For LLMs/MLLMs researchers, the primary value lies in intrinsic interpretability and controllability. Because every decision made by the model is preceded by a rationale grounded in an interpretable, behaviorally validated dimension, the decision processes become directly auditable against known human criteria. Moreover, the proposed framework is domain-agnostic. It enables researchers to convert existing large-scale cognitive science datasets into rich process-level supervision, facilitating the development of controllable, consensus-consistent LLMs/MLLMs.

For neuroscientists and cognitive scientists, the models additionally offer a new class of testable computational models and can be subjects for hypothesis generation about related tasks to reduce the difficulty and cost of data sampling. Our work also encourages relevant researchers to shift their focus from outcome-level studies to a deeper exploration of the underlying cognitive reasoning processes.

F Method Details

F.1 Internal Feature Extraction via Forward Propagation

To map the MLLMs internal representations to human brain activity, we construct a multimodal input using a neutral instruction prompt (`Describe this image`) alongside the visual stimuli and perform forward propagation pass through baseline and aligned MLLMs. Targeting the most representative feature, we extract the hidden states specifically from the `<|vision_end|>` token which is the special token delimiting the end of the visual embedding sequence for Qwen3-VL series models. We extract these states from three specific depths: the last layer, the second-to-last layer, and the last four layers [Devlin et al., 2019]. We prioritize this token over individual spatial image pad tokens or subsequent text tokens for two fundamental reasons: (1) Global Visual Integration: Unlike individual image pad tokens which encode spatially localized features, the `<|vision_end|>` token serves as a global aggregator. Due to the causal attention mechanism of the LLM backbone, this token attends to the entire preceding visual sequence, effectively summarizing the holistic visual semantic content before text begins. (2) Vision-Language Boundary: This token represents the *terminal state* of visual perception and the *initial state* of text. It is the bottleneck where visual features are projected into the semantic manifold, making it the most theoretically plausible candidate for alignment with human cortical representations.

G Data Details

G.1 Statistics

Table 6 summarizes the composition of our hybrid training dataset, reporting sample counts and proportions for cognitive-alignment data and generalization data across single-image and multi-image tasks.

G.2 Cognitive-Alignment Data

O1O task data to align the latent structure of human cognition. $\mathcal{D}_{\text{O1O}}$ is constructed via method 3.2 using prompt 14.

G.3 Generalization Data

We supplement $\mathcal{D}_{\text{O1O}}$ with a diverse set of regularization data $\mathcal{D}_{\text{gen}}$ selected from open-source datasets, categorized into Single-image and multi-image scenarios:

Single-image scenarios. We incorporate COCO [Lin et al., 2014], ALLaVA-4V [Chen et al., 2024], and LLaVAR [Zhang et al., 2024].

Multi-image scenarios. Since the Odd-One-Out task inherently involves comparing multiple images (a triplet), we specifically include multi-image reasoning tasks from M4-Instruct [Li et al., 2025], such as Spot the Difference [Jhamtani and Berg-Kirkpatrick, 2018, Park et al., 2019], Image Edit Instruction [Zhang et al., 2023, Bodur et al., 2024], Visual Storytelling [Ravi et al., 2021,

Table 6: Hybrid dataset composition.

Task	**Dataset**	**Samples**	**Ratio** (%)
Single-image Scenarios			
Caption (13.5K)	COCO	13,501	4.36
ALLaVA-4V (69.98K)	ALLaVA-VFLAN	19,990	6.46
	ALLaVA-LAION	49,990	16.16
LLaVAR (43.17K)	LLaVAR-GPT4	43,167	13.96
Multi-image Scenarios			
Spot the Difference (14.66K)	Spot-the-Diff	10,774	3.48
	CLEVR-Change	3,885	1.26
Image Edit Instruction (23.013K)	MagicBrush	19,557	6.32
	IEdit	3,456	1.12
Visual Storytelling (32.94K)	AESOP	6,915	2.24
	VIST	26,026	8.41
Text-rich VQA (21.39K)	WebQA	9,338	3.02
	TQA	8,249	2.67
	OCR-VQA	1,899	0.61
	DocVQA	1,900	0.61
Low-level Comparison (10.68K)	DreamSim	10,682	3.45
Odd-One-Out Task			
Cognitive-Alignment (80K)	THINGS	80,000	25.86
Total		309329	100.0

Huang et al., 2016], Text-rich VQA, and Low-level Comparison [Fu et al., 2023, Sundaram et al., 2024]. This ensures that the model learns to process multi-image text sequences robustly, rather than overfitting to the specific three-image format of the O1O task.

G.4 Dynamic Stratified Data Sampling

Given the diverse nature of our training mixture (ranging from O1O cognitive data to general data), utilizing a naive data loader could lead to distributional shifts between batches (e.g., a batch dominated solely by one task type). To mitigate this, we implement a dynamic stratified sampling strategy.

Specifically, we define the sampling probability for each task k based on its proportion in the total dataset. During training, for every global step, we construct the mini-batch by sampling data points from each task category proportional to these ratios, constrained by the total global batch size across all GPUs. The samples within each constructed batch are then shuffled using a fixed random seed.

This strategy guarantees that every gradient update vector is computed from a representative distribution of both cognitive alignment signals and general regularization data. By stabilizing the gradient direction and preventing task-specific bias during optimization, this sampling method is helpful for retaining the model's generalization capabilities while effectively instilling the latent structure of human cognition.

H Experimental Details

We provide a comprehensive overview of our experimental setup, including the hyperparameters for Supervised Fine-Tuning (SFT).

H.1 Training Details

Supervised fine-tuning (SFT). As shown in Table 7, the SFT stage utilizes AdamW [Loshchilov and Hutter, 2019] and a cosine learning rate scheduler with a peak learning rate of 1e-5. We total use 6 $\times$ NVIDIA A800 80G GPUs with TRL [von Werra et al., 2020], Qwen-VL utils [QwenLM, 2026], and DeepSpeed [Rajbhandari et al., 2020].

Table 7: SFT params.

Parameter	Value
Optimizer	AdamW
Learning Rate	1e-5
Scheduler	cosine
Weight Decay	0.0
Epochs	1
Warmup Ratio	3%
Batch Size	144
Seed	42

Table 8: Eval params.

Parameter	Value
Engine	vLLM
Precision	BF16
Temperature	0.0
Top-p	1.0
Top-k	-1
Rep. Penalty	1.05
Max Tokens	8192

H.2 Evaluation Details

We conduct a comprehensive evaluation using the vLLM [Kwon et al., 2023]. For evaluation metrics, we adopt traditional exact string matching to accurately assess response correctness about human consistency. Following the guidelines from the official vLLM documentation, we set the temperature to 0.0, top-p to 1.0, and top-k to -1. We apply a repetition penalty of 1.05 and limit the maximum response length to 8192 tokens.

To ensure a strictly fair comparison, we re-evaluate all baselines on our local infrastructure rather than relying on reported metrics. We observe a performance discrepancy between local reproduction and reported results, a known issue in benchmark evaluation [ArdalanM, 2025]. Therefore, we report results for both the Baseline and our CogAlign models under the exact same experimental setting. Specifically, we employ the lmms-eval framework [Zhang et al., 2025b, Li* et al., 2024] and adhere strictly to its default parameters for MMMU and MMBench benchmarks. By using the standard default parameters, we eliminate potential biases arising from hyperparameter tuning.

H.3 Generalization Performance Details

Table 9 reports detailed results on the MMBench(en,test) for the baseline and CogAlign variants of Qwen3-VL-4B, Qwen3-VL-8B, and Qwen3-VL-30B-A3B. In addition to overall scores, it provides breakdowns by general capability and specific task, allowing a finer-grained comparison of performance before and after cognitive alignment.

Table 9: Detailed performance comparison of MMBench (en,test).

Model	Qwen3-VL 4B		8B		30B-A3B	
	Baseline	CogAlign	Baseline	CogAlign	Baseline	CogAlign
A. Overall	82.96	81.39	83.91	83.86	87.44	83.86
B. General Capabilities						
AR	82.99	85.76	80.90	87.85	87.15	87.50
CP	87.58	85.44	89.08	87.15	88.65	86.30
FP-C	82.19	74.09	78.95	74.90	85.43	81.78
FP-S	87.94	87.69	90.20	90.20	93.72	88.69
LR	71.68	67.63	73.99	74.57	79.77	73.41
RR	73.46	74.41	78.67	77.25	81.99	75.36
C. Specific Tasks						
Action Recognition	93.18	90.91	92.05	89.77	93.18	93.18
Attribute Comparison	85.90	80.77	87.18	80.77	88.46	87.18
Attribute Recognition	90.91	91.92	90.91	93.94	94.95	90.91
Celebrity Recognition	85.59	87.29	89.83	88.98	93.22	88.14
Function Reasoning	87.74	87.74	87.74	90.57	91.51	90.57
Future Prediction	75.00	73.61	76.39	77.78	80.56	73.61
Identity Reasoning	96.34	96.34	96.34	97.56	98.78	97.56
Image Emotion	77.11	73.49	74.70	72.29	80.72	75.90
Image Quality	85.88	82.35	90.59	88.24	88.24	88.24
Image Scene	92.25	91.47	95.35	93.80	91.47	89.92
Image Style	83.53	80.00	84.71	80.00	83.53	77.65
Image Topic	96.47	96.47	96.47	97.65	97.65	97.65
Nature Relation	77.78	80.25	85.19	82.72	90.12	75.31
Object Localization	81.73	78.85	84.62	83.65	90.38	81.73
OCR	96.10	94.81	97.40	96.10	97.40	96.10
Physical Prop. Reas.	67.00	75.00	61.00	77.00	73.00	76.00
Physical Relation	51.92	57.69	48.08	51.92	57.69	65.38
Social Relation	83.33	79.49	92.31	88.46	89.74	82.05
Spatial Relationship	66.67	49.38	56.79	53.09	74.07	64.20
Struct. Image-Text	69.31	63.37	72.28	72.28	79.21	73.27

H.4 Necessity of Alignment over Prompting

We use a more lenient parser for d^*-CoT and FS-CoT settings because of their generation failures and manually examine the parsing failures. We find that almost all failures are not caused by minor formatting variations or an overly strict parser. Instead, the model frequently enters a highly repetitive generation loop, repeatedly enumerating irrelevant positive and negative attributes. Consequently, it exhausts the maximum generation length before producing the required closing tag. Thus, such cases represent genuine generation failures rather than correct predictions that are rejected solely because of formatting. We list an example of parsing failures below.

<thinking>
First, let's analyze each object using the provided semantic dimensions.
Object A: A rabbit. This is an animal. It is soft, fluffy, and related to nature. It is not metallic, not artificial, not food-related, not human-related, not tool-related, not plant-related, not house-related, not

transportation-related, not circular, not flat, not long or thin, not pointed, not flying, not colorful, not white, not green, not yellow, not orange, not ....

H.5 Necessity of Dimension-Grounded Rationales

To isolate the exact drivers of our framework alignment gains, we conduct ablations on the content of the supervision signal. First, replacing the rationale with a sparse, single-word dimension label severely degrades performance as shown in Table 10. Second, fine-tuning on choice-only and generic CoT rationales fails to match the performance of our dimension-grounded rationales in Table 11, confirming that the gains stem specifically from the injection of human cognitive priors rather than task-specific fine-tuning.

Table 10: Ablation on supervision format during fine-tuning Qwen3-VL-4B-Instruct.

Metric	Baseline (CoT)	Dimension-only	CogAlign (Ours)
Accuracy	53.37	42.81	**56.81**
Choice-based RSA	81.79	44.01	**85.01**

Table 11: Comparison of Choice-only, Generic CoT, and ours supervision signals.

Model	Baseline (CoT)	Generic CoT	Choice-only	CogAlign (ours)
Qwen2.5-VL-7B-Instruct	56.63	—	56.21	**59.29**
Qwen3-VL-4B-Instruct	53.37	47.57	—	**56.81**

H.6 Neuroscientific Analysis Details

Positive RSA shifts are observed across the ventral occipito-temporal cortex, encompassing category-selective regions such as the EBA, FFA, and PPA. This indicates that the aligned MLLMs representational geometry better captures the high-level semantic organization for object and category discrimination. Because these alignment gains are localized to high-level visual cortices, they reflect an abstraction of category- and shape-relevant semantic structures, rather than mere low-level retinotopic or pixel-wise feature matching.

Furthermore, we observe substantial alignment shifts extending into the IPS and RSC. Given the triplet comparison nature of the O1O task, the engagement of the IPS aligns with its established role in the dorsal attention and multiple-demand networks, facilitating top-down attentional selection and relational comparison. Concurrently, shifts within the RSC likely indicate enhanced alignment with memory-linked contextual and scene-associative representations. Taken together, these neural effects demonstrate that our framework drives the MLLMs to capture behaviorally relevant, higher-order similarity structures. Beyond behavioral task performance, these results provide a convergent, neural-level validation: anchoring MLLMs to behavior-derived semantic dimensions explicitly steers their internal representations toward human-like cognitive geometry.

H.7 Similarity of Behavioral Representational Geometry Details

To assess whether the global decision-making structure of the models aligns with human cognition, we perform RSA based on the behavioral choices in the O1O task. This section details the construction of the choice-based RSMs and the subsequent calculation of the geometry similarity scores.

Choice-based RSM. We evaluate the representational geometry on a subset of $N = 48$ objects, yielding an $N \times N$ matrix for human and evaluating models (Baseline, CogAlign, and Centaur). We initialize an all-zero similarity matrix $S \in \mathbb{R}^{N \times N}$.

The evaluation dataset consists of approximately 91.56K triplets drawn from this 48-object subset. For each triplet iteration $t = \{o_i, o_j, o_k\}$, the models are tasked with selecting the odd-one-out object using CoT. Let the chosen odd-one-out be $o_k \in t$. The remaining two objects, denoted as (o_i, o_j), are deemed the most similar pair within the context of this triplet. We then update the similarity matrix by incrementing the co-occurrence counts for this pair (i, j) and (j, i):

$$S_{i,j} \leftarrow S_{i,j} + 1 \quad \text{and} \quad S_{j,i} \leftarrow S_{j,i} + 1$$

This process is repeated across all triplets in the subset. The resulting matrix S captures the global behavioral representational geometry of the models. Using this method, we derive the human RSM (S_{human}) and model RSM (S_{model}).

Quantifying representational geometry similarity. We extract the upper triangular portion of their respective symmetric matrices, excluding the main diagonal. Let $\mathbf{v}_{\text{human}} \in \mathbb{R}^D$ and $\mathbf{v}_{\text{model}} \in \mathbb{R}^D$ be the flattened 1D vectors derived from the upper triangles of S_{human} and S_{model} respectively, where $D = N(N-1)/2$. The behavioral representational geometry similarity is quantified using the Pearson correlation:

$$\text{Similarity Score} = \rho(\mathbf{v}_{\text{model}}, \mathbf{v}_{\text{human}}) \tag{14}$$

where $\rho(\cdot, \cdot)$ is the Pearson correlation coefficient across the D feature dimensions.

I Behavior-Derived 66 Semantic Dimensions

As shown in Table 12, we provide the saturated mental representations of natural objects underlying human similarity judgements. They are represented by **66** semantic dimension concepts.

Table 12: Overview of 66 semantic dimension concepts.

Axis	Semantic Concept	Axis	Semantic Concept
0	metallic or artificial	1	food-related
2	animal-related	3	textile
4	plant-related	5	house-related or furnishing-related
6	valuable or precious	7	transportation-related or movement-related
8	body-related or people-related	9	wood-related or brown
10	electronics or technology	11	colorful or playful
12	outdoors	13	circular or round
14	paper-related or flat	15	sports-related or playing-related
16	tools or elongated	17	fluid-related or drink-related
18	water-related	19	oriented or many things
20	decay-related or grainy	21	white
22	coarse pattern or many things	23	red
24	long or thin	25	weapon-related or danger-related
26	black	27	household
28	feminine (stereotypical)	29	body part-related
30	tubular	31	music-related, hearing-related, or hobby-related
32	grid-related or grating-related	33	repetitive or spiky
34	construction-related or craftsman-related	35	spherical or voluminous
36	string-related or stringy	37	seating-related, standing-related, or lying-related
38	flying-related or sky-related	39	disgusting or slimy
40	elliptical or curved	41	sand-colored
42	green	43	bathroom-related or wetness-related
44	yellow	45	heat-related or light-related
46	beams-related or mesh-related	47	foot-related or walking-related
48	box-related or container	49	stick-shaped or cylindrical
50	head-related	51	upright, elongated, or voluminous
52	pointed or spiky	53	child-related or cute
54	farm-related or historical	55	seeing-related, small, or round
56	medicine-related	57	dessert-related
58	orange	59	thin or flat
60	cylindrical, conical, or cushioning	61	coldness-related or winter-related
62	measurement-related or numbers-related	63	fluffy or soft
64	masculine (stereotypical)	65	fine-grained pattern

J Prompts

In this section, we provide the full prompts used in our methods.

J.1 Dimension-Guided Reasoning Enhancement Prompt

Template

You're a knowledgeable expert. Given three objects **[triplet names]**. Here are their brief descriptions below.
[triplet captions]
With the above descriptions, you understand three objects **[triplet names]** simply. Next, you need to incorporate the vast amount of knowledge you have to understand them further.
You need to think carefully and analyse in terms of **[dimension name]**. Next, you need to give your analysis.
Here are some criteria you need to follow:
1.You only need to give your analysis. You must keep words in 200 words. You don't need to output the analysis process and how you analyse.
2.You need to analyse **[triplet names]** based on your vast knowledge, known information and analytical perspective which is **[dimension name]**.
3.Your analysis must be objective, reasonable, fair and logical.
4.If there is not much relationship between the angle of analysis, **[dimension name]** and one or more of **[triplet names]**, please do not carry out a forced analysis because forcing the analysis in terms of **[dimension name]** is not permissible to ensure your analysis is reasonable.
5.You must strictly follow the below response template to output with your analysis.
Here is your response template, You have to follow the template output given below
Analysis: your analysis

Example

triplet names: strawberry, pepper mill, and flask.
triplet captions: strawberry: sweet fleshy red fruit; pepper mill: a mill for grinding pepper; flask: bottle that has a narrow neck.
dimension name: plant related.
Output: Analysis: The strawberry is a fruit that develops from the flower of the strawberry plant (Fragaria $\times$ ananassa), which is a member of the rose family (Rosaceae). It is characterized by its sweet, fleshy, and red flesh, and it is widely cultivated for its edible fruit. The plant itself is a perennial herb with runners that allow it to spread and propagate.
The pepper mill, while not directly related to plants, is used to grind peppercorns, which are the dried berries of the Piper nigrum plant, a flowering vine in the family Piperaceae. Peppercorns are harvested when they are fully mature and then dried, turning them into the familiar black, white, or green peppercorns.
The flask, being a container, does not have a direct botanical connection. However, flasks are often used in the processing and storage of plant-derived substances, such as essential oils, tinctures, and extracts, which are commonly used in various industries including pharmaceuticals, cosmetics, and food.
In summary, the strawberry is directly derived from a plant, the pepper mill is indirectly related through its use with peppercorns, and the flask can be associated with the storage and processing of plant materials.

Prompt 13: Prompt for generating enhanced dimension-guided reasoning on triplet-level using LLM.

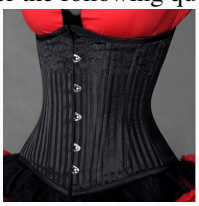
J.2 O1O Task Training Prompt

Template

System:
Here are three important notes:
1.Each image corresponds to a single object.

2.Focus only on the object itself and ignore any background elements.
3.Use reasonable reasoning and answer the question by thinking step by step.
Answer the following question:
User:
[triplet visual images]
Identify the object that is most different from the rest.
Options:
(A) Object A
(B) Object B
(C) Object C
Let's analyze each object.
Assistant:
Object A: [object A caption]
Object B: [object B caption]
Object C: [object C caption]
[dimension-guided reasoning]
It is clear that the most different object is [odd-one-out]. The answer is [right option].

Example

System:
Here are three important notes:
1.Each image corresponds to a single object.
2.Focus only on the object itself and ignore any background elements.
3.Use reasonable reasoning and answer the question by thinking step by step.
Answer the following question:
User:   

Identify the object that is most different from the rest.
Options:
(A) Object A
(B) Object B
(C) Object C
Let's analyze each object.
Assistant:
Object A: corset which is a closely fitted, typically sleeveless garment worn by women to provide support and shape to the torso, often extending from the bust to the hips, and usually fastened with laces or other closures.
Object B: bacon which is a cured meat from the back and sides of a hog, typically salted and dried or smoked, then sliced thin and often fried until crispy, serving as a popular breakfast food or ingredient in various dishes.
Object C: piggy bank which is a traditional children's savings container, typically made of ceramic or plastic, shaped like a pig, used to store coins and teach kids about saving money, often featuring a slot on top for depositing coins and a removable plug for withdrawing the accumulated savings.
Bacon, as a food item, is a cured meat product derived from the belly or side cuts of pork, often smoked and fried, serving as a staple in many cuisines. It is rich in flavor and fat, making it a popular ingredient in various dishes. The piggy bank, while not directly related to food, can symbolically represent savings that might be used to purchase food items, including bacon. However, it has no intrinsic connection to culinary practices or nutritional value. The corset, being a type of clothing, has no direct relationship to food. Historically, corsets were designed for shaping the body and have no functional or symbolic link to food preparation, consumption, or culture. Therefore, while bacon is inherently food-related, the piggy bank and corset do not have significant connections to the food domain.

It is clear that the most different object is **corset**. The answer is A.

Prompt 14: Prompt for O1O task training data to align the latent structure of human cognition.

J.3 O1O Direct Inference Prompt

Input
System:
Here are three important notes:
1.Each image corresponds to a single object.
2.Focus only on the object itself and ignore any background elements.
3.Answer the question with correct option directly (e.g., A, B, or C).
Answer the following question:
User:
[triplet visual images]
Which object is the odd one out?
Options:
(A) Object A
(B) Object B
(C) Object C

Prompt 15: Prompt for O1O direct inference. MLLMs only answer with predicted option (A/B/C).

J.4 O1O CoT Inference Prompt

Standard CoT
System:
Here are three important notes:
1.Each image corresponds to a single object.
2.Focus only on the object itself and ignore any background elements.
3.Use reasonable reasoning and answer the question by thinking step by step.
Answer the following question:
User:
[triplet visual images]
Which object is the odd one out?
Options:
(A) Object A
(B) Object B
(C) Object C
At the very last line of your response, must put correct answer: your chosen option. Do not add anything after it.
Let's analyze each object.

Prompt 16: Prompt for O1O standard CoT inference. MLLMs need to think step-by-step and give final choice. This prompt will be used by baseline and CogAlign for CoT, S-CoT metrics.

d^* CoT
System:
Here are two important notes:
1.Each image corresponds to a single object.
2.Focus only on the object itself and ignore any background elements.
Answer the following question:
User:
[triplet visual images]
Which object is the odd one out?
Options:

(A) Object A
(B) Object B
(C) Object C.
Thinking Perspective:
Before providing your final choice, you must conduct deep, step by step thinking from the perspective of **[dimension name]**.
CRITICAL Output Format:
You MUST strictly adhere to the exact structure below. Do not add any extra words outside of this template.
<thinking>
[Write your detailed step-by-step reasoning process here]
</thinking>
Therefore, the answer is <choice>[A, B, or C]</choice>

Prompt 17: Prompt for O1O d^* dimension-guided CoT inference. MLLMs need to think step-by-step and give final choice under the guidance of d^* semantic dimension. This prompt will be used by baseline models for d^*-CoT metric.

FS CoT

System:
Here are two important notes:
1.Each image corresponds to a single object.
2.Focus only on the object itself and ignore any background elements.
Answer the following question:
User:
[triplet visual images]
Which object is the odd one out?
Options:
(A) Object A
(B) Object B
(C) Object C.
Available Semantic Dimensions:
dim1: metallic or artificial
dim2: food-related
dim3: animal-related
dim4: textile
dim5: plant-related
dim6: house-related or furnishing-related
dim7: valuable or precious
dim8: transportation-related or movement-related
dim9: body-related or people-related
dim10: wood-related or brown
dim11: electronics or technology
dim12: colorful or playful
dim13: outdoors
dim14: circular or round
dim15: paper-related or flat
dim16: sports-related or playing-related
dim17: tools or elongated
dim18: fluid-related or drink-related
dim19: water-related
dim20: oriented or many things
dim21: decay-related or grainy
dim22: white
dim23: coarse pattern or many things
dim24: red
dim25: long or thin
dim26: weapon-related or danger-related
dim27: black
dim28: household
dim29: feminine (stereotypical)

dim30: body part-related
dim31: tubular
dim32: music-related, hearing-related, or hobby-related
dim33: grid-related or grating-related
dim34: repetitive or spiky
dim35: construction-related or craftsman-related
dim36: spherical or voluminous
dim37: string-related or stringy
dim38: seating-related, standing-related, or lying-related
dim39: flying-related or sky-related
dim40: disgusting or slimy
dim41: elliptical or curved
dim42: sand-colored
dim43: green
dim44: bathroom-related or wetness-related
dim45: yellow
dim46: heat-related or light-related
dim47: beams-related or mesh-related
dim48: foot-related or walking-related
dim49: box-related or container
dim50: stick-shaped or cylindrical
dim51: head-related
dim52: upright, elongated, or voluminous
dim53: pointed or spiky
dim54: child-related or cute
dim55: farm-related or historical
dim56: seeing-related, small, or round
dim57: medicine-related
dim58: dessert-related
dim59: orange
dim60: thin or flat
dim61: cylindrical, conical, or cushioning
dim62: coldness-related or winter-related
dim63: measurement-related or numbers-related
dim64: fluffy or soft
dim65: masculine (stereotypical)
dim66: fine-grained pattern
Thinking Perspective:
Before providing your final choice, you must conduct deep, step by step thinking from the perspective of one or multiple dimensions chosen from the **Available Semantic Dimensions** list above.
CRITICAL Output Format:
You MUST strictly adhere to the exact structure below. Do not add any extra words outside of this template.
<thinking>
[Write your detailed step-by-step reasoning process here]
</thinking>
Therefore, the answer is <choice>[A, B, or C]</choice>

Prompt 18: Prompt for O1O 66 dimensions CoT inference. MLLMs need to think step-by-step from one or more semantic dimension(s) and give final choice. This prompt will be used by baseline models for FS-CoT metric.

K Impact Statement

This work presents research on aligning Multimodal Large Language Models (MLLMs) with latent structure of human cognition. The primary positive impact is advancing our understanding of MLLMs by thinking process before decision-making, providing a data-driven framework to bridge the representational gap between models and human similarity judgements, making MLLMs more human-like. However, because our framework relies on crowdsourced behavioral data provided by Hebart et al., it carries the inherent risk of internalizing and transferring implicit human cognitive

biases or stereotypes present in the dataset. To mitigate this, we advise future research to audit behavior-derived datasets for cultural and demographic biases prior to alignment.

L Dataset Licenses

We use the M4-Instruct dataset [Li et al., 2025], THINGS behavioral dataset [Hebart et al., 2020, 2023], and Natural Scenes Dataset (NSD) [Allen et al., 2022] following their respective license agreements. Both datasets are accessed through their official releases and use exclusively for academic research purposes.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We summarize the main contributions in the Introduction section 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in the Limitations section of the Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We disclose data preparation, training and evaluation settings, feature extraction, and prompts in Section 4 and Appendix F.1, G, H, and J. The released code and rationale data provide the configurations and procedures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and data is available at `https://github.com/ening-open/cogalign`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify the training and test details in the Section 4 and Appendix G, and H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide confidence intervals and statistical significance tests in the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide detailed training and evaluation settings in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research uses publicly available behavioral datasets and fMRI datasets collated under the original studies' ethical approvals.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts in Appendix K.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited all necessary papers and the license information and usage details of the datasets are provided in Appendix L.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code and related data are documented at `https://github.com/ening-open/cogalign`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are employed to convert the most salient decision dimension into a dimension-grounded linguistic rationale, which may impact the alignment performance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.