# THE MODEL'S OWN TAXONOMY: HOW MLLMS ORGANIZE HUMAN COGNITION TASKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Human cognitive theories provide taxonomies of abilities involved in perception and reasoning, but how closely these taxonomies characterize multimodal large language model (MLLM) reasoning remains understudied. To examine this correspondence, we cluster examples using model hidden-state representations and compare the resulting clusters with human cognitive categories. Across five MLLMs, these clusters reveal three relationships with the human taxonomy, including close correspondence, subdivision within categories, and regrouping across category boundaries. To test whether these distinctions matter for reasoning, we learn a shared attention-head intervention for each human cognitive category or model cluster and compare its effects on accuracy. Cluster-guided interventions outperform those guided by human cognitive categories and improve held-out test accuracy by 3.75-7.98% over the unmodified models. Together, these findings suggest that human cognitive categories capture part of the organization of model reasoning, while model clusters identify additional distinctions that support effective reasoning interventions.

## 1 INTRODUCTION

Building artificial systems that learn and reason as flexibly as humans is a central goal of artificial intelligence (Tenenbaum et al., 2011; Lake et al., 2015; Lake & Baroni, 2023). Benchmarks help assess progress toward this goal by evaluating models across diverse tasks, such as object recognition and geometric reasoning (Liu et al., 2024; Yue et al., 2024; Lu et al., 2024; Fu et al., 2025).

Developmental cognitive science provides a framework for organizing these task evaluations around human abilities in domains such as objects, number, and space (Piaget & Inhelder, 1969; Hermer & Spelke, 1994; Spelke & Kinzler, 2007). For multimodal large language models (MLLMs), CoreCognition draws on this framework to assess model strengths and weaknesses across twelve categories grounded in developmental psychology (Li et al., 2025). However, it remains unclear to what extent model reasoning follows human cognitive distinctions and in what ways its organization differs from the human taxonomy.

To examine this correspondence from the model's own perspective, we cluster examples using hidden-state representations, which encode task-relevant information about how the model processes and represents each example (Wu & Varshney, 2025). We then compare these model clusters with human cognitive categories to examine where their organization agrees and differs. To test whether these distinctions matter for reasoning, we use model interventions, e.g., attention-head interventions (Nam et al., 2025), tailored to each human cognitive category or model cluster. Comparing their effects on accuracy provides functional evidence about whether examples within the same category or cluster share reasoning patterns. In this way, we examine how far human cognitive categories characterize model reasoning and whether model clusters capture additional distinctions that affect it.

Across five MLLMs, our representational analyses reveal three relationships between human cognitive categories and model clusters. Model clusters can (1) closely correspond to individual human cognitive categories, (2) subdivide a category into multiple clusters, or (3) regroup examples across categories. These patterns show that model representations align with some human cognitive distinctions while also exhibiting an organization that the human taxonomy does not fully capture. The intervention results further suggest that model clusters align more closely with model reasoning



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

patterns than human cognitive categories. Specifically, cluster-guided attention-head interventions improve test accuracy by 3.75-7.98% over the unmodified models and outperform interventions based on human cognitive categories by 1.55-3.08%. On Qwen3-VL, these interventions also outperform the two input-dependent steering baselines evaluated in our experiments. Together, these findings suggest that human cognitive categories capture some aspects of model reasoning, while model clusters reveal additional distinctions that support reasoning adaptation.

Our contributions are summarized as follows:

- We compare a developmental cognitive taxonomy with clusters derived from the representations of five MLLMs, revealing close correspondence, subdivision, and regrouping between human cognitive categories and model clusters.
- We use attention-head interventions to examine whether examples within each human cognitive category or model cluster share reasoning patterns, providing functional evidence for cluster-specific patterns. Our method improves test accuracy by 3.75-7.98% over the unmodified models across all five MLLMs and outperforms the two input-dependent steering methods evaluated on Qwen3-VL.

## 2 Related Work

**Cognitive Evaluation of MLLMs.** Multimodal benchmarks evaluate model abilities across diverse tasks (Fu et al., 2025; Liu et al., 2024; Yue et al., 2024), including mathematical problem solving (Lu et al., 2024) and perspective taking (Tang et al., 2025). Cognitive evaluations organize this assessment around human abilities, drawing on psychological experiments (Binz & Schulz, 2023) and developmental data (Tan et al., 2024). For example, CoreCognition examines whether MLLMs possess foundational human cognitive abilities (Li et al., 2025). However, it remains unclear how closely model reasoning follows the human cognitive categories used in these evaluations. We examine this relationship by comparing human cognitive categories with model clusters and testing their functional relevance through shared attention-head interventions.

**Attention Head Functions and Interventions.** Prior work suggests that groups of attention heads can support different functions, including in-context learning (Todd et al., 2024), information retrieval from long contexts (Wu et al., 2025), and logical reasoning (Ma et al., 2025). Head masks specify groups of attention heads for intervention. Consistent performance changes across examples under a shared mask could provide functional evidence of similar reasoning patterns among those examples (Ma et al., 2025; Nam et al., 2025). We therefore learn a shared head mask for each human cognitive category or model cluster and compare the effects of these masks on accuracy.

**Input-Dependent Adaptation.** Input-dependent adaptation uses input representations to apply interventions (Lee et al., 2025; Wang et al., 2025b; 2026). For example, RISER uses a learned router to compose reusable reasoning vectors for each input (Ye et al., 2026). These methods adapt interventions to individual inputs, but leave unclear how the model reasoning patterns underlying effective adaptation relate to human cognitive categories. We investigate this relationship by comparing interventions guided by human cognitive categories and clusters of the target MLLM's hidden states.

## 3 Method

Our framework connects cognitive comparison with reasoning adaptation. We first identify clusters from model representations and compare them with human cognitive categories. We then optimize attention-head masks for each category or cluster and use their effects on accuracy to examine whether the grouped examples share similar reasoning patterns.

### 3.1 Data and Notation

Let $\mathcal{D} = \{(x_i, y_i, c_i)\}_{i=1}^{N}$ contain multimodal examples. Each $x_i$ includes images, a question, and answer options. Here $y_i$ is the correct answer, and $c_i \in \{1, \ldots, C\}$ denotes a human cognitive category, with $C = 12$ in CoreCognition (Li et al., 2025). We split $\mathcal{D}$ into disjoint training, validation, and test sets, denoted by $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{test}}$, respectively.

### 3.2 REPRESENTATIONAL COMPARISON

We compare human cognitive categories and model clusters on $\mathcal{D}_{\text{train+val}} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$.

The dataset's cognitive labels define $\mathcal{H} = \{H_j\}_{j=1}^{C}$, where each category contains the examples with label $j$,

$$H_j = \{(x_i, y_i, c_i) \in \mathcal{D}_{\text{train+val}} \mid c_i = j\}. \tag{1}$$

To characterize how the model organizes examples, we use the hidden states of $x_i$, which reflect how the model internally represents the example (Wu & Varshney, 2025). Specifically, for model $m$, let $\mathbf{h}^{(m)}(x_i) \in \mathbb{R}^{d_m}$ denote the normalized hidden-state representation at the last prompt token, where $d_m$ is the hidden dimension. This state provides a contextualized representation of $x_i$, incorporating its visual and textual information. KMeans fits $K$ centroids $\boldsymbol{\mu}_k^{(m)} \in \mathbb{R}^{d_m}$ to the hidden-state representations. Given these centroids, the nearest-centroid rule assigns each example $x_i$ to a cluster through

$$z^{(m)}(x_i) = \arg\min_{k \in \{1,\ldots,K\}} \left\| \mathbf{h}^{(m)}(x_i) - \boldsymbol{\mu}_k^{(m)} \right\|_2^2. \tag{2}$$

These cluster assignments define $\mathcal{M}^{(m)} = \{M_k^{(m)}\}_{k=1}^{K}$, where

$$M_k^{(m)} = \{(x_i, y_i, c_i) \in \mathcal{D}_{\text{train+val}} \mid z^{(m)}(x_i) = k\}. \tag{3}$$

With $\mathcal{H}$ as a cognitive reference, we examine whether these clusters recover human cognitive categories, subdivide them, or connect examples across their boundaries. Comparisons of $\mathcal{M}^{(m)}$ across models then test which relationships recur across architectures and which depend on the model.

### 3.3 FUNCTIONAL COMPARISON

To examine whether examples within each human cognitive category $H_j$ or model cluster $M_k^{(m)}$ share similar reasoning patterns, we apply a shared attention-head mask and measure the resulting changes in accuracy (Ma et al., 2025; Nam et al., 2025). These changes provide functional evidence about similarities in how the model processes the examples.

Following Causal Head Gating (CHG) (Nam et al., 2025), we parameterize the shared mask as a vector $\mathbf{g}$ with one coefficient $g_h \in [0, 1]$ per head. Here $h$ indexes all attention heads. The mask scales each head output as

$$\widetilde{\mathbf{o}}_h(x_i) = g_h \mathbf{o}_h(x_i). \tag{4}$$

Here $\mathbf{o}_h(x_i)$ is the original head output and $\widetilde{\mathbf{o}}_h(x_i)$ is its masked output.

To identify head contributions that support examples within the same category or cluster, we optimize a shared mask on each $H_j \cap \mathcal{D}_{\text{train}}$ or $M_k^{(m)} \cap \mathcal{D}_{\text{train}}$. For the multiple-choice dataset (Li et al., 2025), we set the task loss $\ell(\mathbf{g})$ as cross-entropy loss over candidate-option scores, with the correct option as the target. The objective combines this task loss with sparsity regularization $R(\mathbf{g})$ (Nam et al., 2025),

$$\mathbf{g}^\star = \arg\min_{\mathbf{g}} \left[ \ell(\mathbf{g}) + \lambda R(\mathbf{g}) \right]. \tag{5}$$

Here $\lambda \geq 0$ controls the regularization strength. Optimization updates only the masks and keeps the backbone weights fixed. The resulting masks are denoted by $\mathbf{g}^\star_{H_j}$ and $\mathbf{g}^\star_{M_k^{(m)}}$, with the model index implicit for masks learned for human cognitive categories. Appendix B.2 provides the parameterization and optimization details.

We then apply the mask learned for each human cognitive category or model cluster to its examples in $\mathcal{D}_{\text{val}}$ and assess whether examples within the same category or cluster share reasoning patterns.

We further examine whether model clusters can guide interventions that improve model reasoning. To this end, we evaluate cluster-guided interventions on $\mathcal{D}_{\text{test}}$ that contribute to neither clustering nor mask optimization. For each $x_i$, its representation from the unmodified model determines the nearest cluster $k = z^{(m)}(x_i)$ through Equation (2). The corresponding mask $\mathbf{g}^\star_{M_k^{(m)}}$ then adjusts the head contributions during inference.

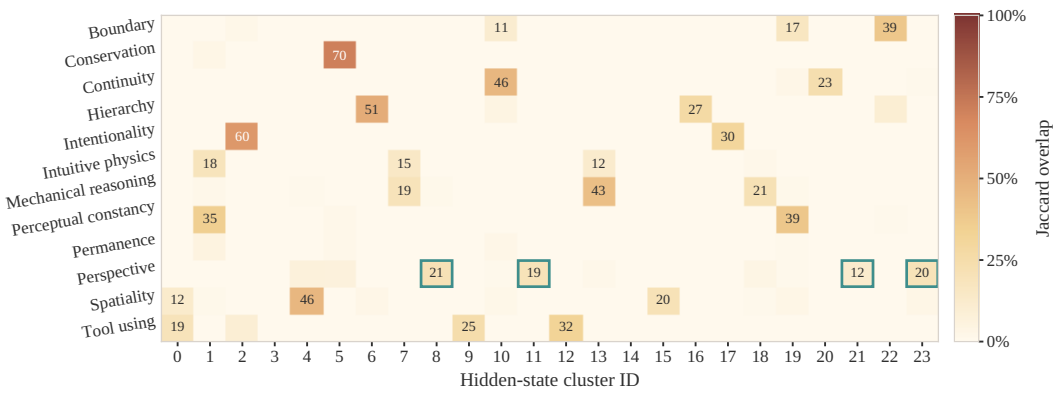



Figure 1: Jaccard overlap between human cognitive categories $H_j$ and Qwen3-VL clusters $M_k^{(m)}$.

## 4 Experiments

Our experiments examine how closely human cognitive categories correspond to model representations and whether examples within each category or model cluster share reasoning patterns. We further evaluate whether cluster-guided interventions improve model reasoning and which factors influence their effectiveness.

### 4.1 Experimental Setup

**Models.** We study five MLLMs from three model families, namely Qwen3-VL-8B-Instruct, Qwen3.5-9B, Gemma 3 12B IT, Gemma 4 12B IT, and GLM-4.6V-Flash. We abbreviate them as Qwen3-VL, Qwen3.5, Gemma 3, Gemma 4, and GLM-4.6V. Our main comparison uses twelve human cognitive categories in $\mathcal{H}$ and $K = 24$ model clusters in each $\mathcal{M}^{(m)}$. We also examine whether the target model's own hidden states better capture its reasoning patterns than representations from an external embedding model, i.e., Qwen3-VL-Embedding-2B (Li et al., 2026). Appendix B specifies layers of the hidden states, mask dimensions, and training settings.

**Data.** We extend the CoreCognition dataset (Li et al., 2025) according to its cognitive prototypes, yielding 3,294 non-video multimodal examples (1,377 in $\mathcal{D}_{\text{train}}$, 917 in $\mathcal{D}_{\text{val}}$, and 1,000 in $\mathcal{D}_{\text{test}}$) and 172 video examples. Clustering uses $\mathcal{D}_{\text{train+val}}$ together with the 172 video examples. Representational comparisons and intervention experiments use non-video examples, keeping the input modality consistent across these analyses. Appendix A details source datasets and construction procedures.

**Evaluation.** To assess how closely model clusters follow human cognitive distinctions, we measure overall agreement between $\mathcal{H}$ and $\mathcal{M}^{(m)}$ using normalized mutual information (NMI) (Strehl & Ghosh, 2002) and adjusted mutual information (AMI) (Vinh et al., 2009). Jaccard overlap complements these measures by quantifying shared example membership between individual human cognitive categories and model clusters. To examine whether models organize the same examples similarly, we compare their clusters using the adjusted Rand index (ARI) (Hubert & Arabie, 1985) and NMI.

We evaluate intervention effects using accuracy. To reduce sensitivity to answer positions, we cyclically rotate the options so that the correct answer occupies each position once and report accuracy across all variants. For each variant, we select the option with the lowest mean token negative log-likelihood.

### 4.2 Cognitive Correspondence and Reasoning Adaptation

**Partial Alignment with Human Cognitive Categories.** We first examine whether human cognitive categories capture the similarities among examples that emerge in model representations.

Table 1 compares $\mathcal{H}$ with each $\mathcal{M}^{(m)}$ on $\mathcal{D}_{\text{train+val}}$. NMI and AMI show partial agreement across all five models. The mean best Jaccard score, which averages each human cognitive category's strongest overlap with a model cluster, also indicates that a single cluster does not generally recover a full category.

Table 1: Agreement between the human cognitive taxonomy $\mathcal{H}$ and each model's clusters $\mathcal{M}^{(m)}$.

| Model | NMI | AMI | Mean best Jaccard |
|---|---|---|---|
| Qwen3-VL | 0.6651 | 0.6582 | 0.3914 |
| Qwen3.5 | 0.6543 | 0.6471 | 0.4008 |
| Gemma 3 | 0.6197 | 0.6117 | 0.3444 |
| Gemma 4 | 0.6291 | 0.6215 | 0.4010 |
| GLM-4.6V | 0.5839 | 0.5752 | 0.3288 |

Figure 1 illustrates three relationships between human cognitive categories and Qwen3-VL clusters. (1) **A human cognitive category can closely correspond to a single model cluster**, as *Conservation* does with $M_5^{(m)}$. (2) **A human cognitive category can subdivide into several model clusters**. For example, *Perspective* overlaps with $M_k^{(m)}$ for $k \in \{8, 11, 21, 23\}$, while *Hierarchy* has its strongest overlaps with $M_6^{(m)}$ and $M_{16}^{(m)}$. (3) **A model cluster can regroup examples across human cognitive categories**, as $M_7^{(m)}$ contains examples from both *Intuitive Physics* and *Mechanical Reasoning*. These relationships suggest that models partly organize examples according to human cognitive distinctions, while also exhibiting their own organization that the human cognitive taxonomy does not fully capture. Appendix C.1 presents the corresponding analyses for the other models.

Figure 2 uses t-SNE to visualize close correspondence, subdivision, and regrouping between human cognitive categories and model clusters. Figure 11 in the appendix provides concrete examples of subdivision within *Perspective*. Cluster $M_8^{(m)}$ groups questions about projecting 3D shapes into 2D views, whereas $M_{11}^{(m)}$ groups questions about transforming spatial relations between reference frames. This difference between geometric projection and reference-frame transformation may explain why the model separates examples within the same human cognitive category. Figure 12 in the appendix provides examples of regrouping across categories, with $M_7^{(m)}$ containing both *Intuitive Physics* questions about billiard trajectories and *Mechanical Reasoning* questions about gear rotation. Both require inferring motion from spatial arrangements and physical constraints, which may explain their grouping within the same cluster. These examples suggest that similarities and differences in the required reasoning operations help explain where model clusters follow or depart from human cognitive categories.

**Model Clusters Guide Effective Reasoning Interventions.** The partial representational alignment raises the question of whether model clusters capture shared reasoning patterns more effectively than human cognitive categories. We optimize one attention-head mask for each $H_j$ or $M_k^{(m)}$ on $\mathcal{D}_{\text{train}}$ and evaluate it on $\mathcal{D}_{\text{val}}$. As shown in Table 2, across all five models, both approaches improve accuracy over the unmodified model, and cluster-guided interventions consistently outperform those guided by human cognitive categories. These results suggest that model clusters better reflect each model's own cognitive organization than human cognitive categories do.

Beyond characterizing the model's cognitive organization, these clusters provide a basis for selecting attention-head interventions to improve accuracy. We evaluate these masks on $\mathcal{D}_{\text{test}}$, which we exclude from both clustering and mask optimization. For each $x_i$, we use its representation from the unmodified model to identify the nearest cluster centroid and select the corresponding mask. As shown in Table 3, cluster-guided interventions outperform both the unmodified model and interventions guided by human cognitive categories across all five models.

In addition, we compare our method with two input-dependent steering methods on Qwen3-VL. L2S (Parekh et al., 2025) predicts an input-specific steering vector, while IDEEA (Wang et al., 2026) selects head-level steering directions based on activation-cluster matching. We adapt both methods to our multiple-choice task and fit their components using $\mathcal{D}_{\text{train}}$. As shown in Table 4, our method achieves higher accuracy than both adapted baselines on $\mathcal{D}_{\text{test}}$. These results support

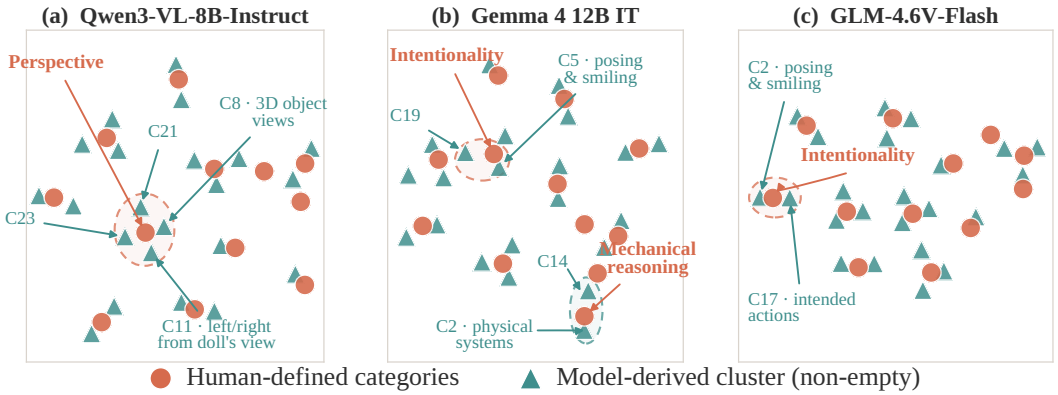



Figure 2: Human cognitive categories $H_j$ and model clusters $M_k^{(m)}$ visualized by t-SNE of binary example-membership vectors with Jaccard distance. Annotations illustrate task forms within selected model clusters. Quantitative comparisons use the original overlaps, not distances in the projection.

Table 2: Accuracy (%) on $\mathcal{D}_{\text{val}}$ with no intervention (Original) or masks for human cognitive categories ($\mathcal{H}$) and model clusters ($\mathcal{M}^{(m)}$).

| Model | Original | $\mathcal{H}$ | $\mathcal{M}^{(m)}$ |
|---|---|---|---|
| Qwen3-VL | 64.16 | 69.38 | **72.56** |
| Qwen3.5 | 66.70 | 73.13 | **73.49** |
| Gemma 3 | 59.17 | 67.47 | **69.51** |
| Gemma 4 | 65.53 | 68.51 | **71.15** |
| GLM-4.6V | 64.14 | 72.86 | **74.80** |

Table 3: Accuracy (%) on $\mathcal{D}_{\text{test}}$ with no intervention (Original) or masks for human cognitive categories ($\mathcal{H}$) and model clusters ($\mathcal{M}^{(m)}$).

| Model | Original | $\mathcal{H}$ | $\mathcal{M}^{(m)}$ |
|---|---|---|---|
| Qwen3-VL | 71.55 | 72.22 | **75.30** |
| Qwen3.5 | 74.48 | 76.57 | **78.96** |
| Gemma 3 | 60.27 | 66.27 | **68.25** |
| Gemma 4 | 70.34 | 73.78 | **75.33** |
| GLM-4.6V | 68.91 | 74.13 | **76.51** |

the effectiveness of cluster-guided attention-head interventions for reasoning adaptation. Jointly optimizing a mask over examples within a cluster may capture shared reasoning patterns and reduce sensitivity to biases specific to individual examples. Appendix B.4 describes the implementation details.

Table 4: Qwen3-VL accuracy on $\mathcal{D}_{\text{test}}$ (%).

| Method | Original | L2S (Parekh et al., 2025) | IDEEA (Wang et al., 2026) | Ours |
|---|---|---|---|---|
| Accuracy (%) | 71.55 | 74.92 | 71.68 | **75.30** |

### 4.3 Shared and Model-Specific Cognitive Organization

Beyond comparing models with human cognition, we examine which aspects of cognitive organization different MLLMs share and where they differ. To this end, each model independently clusters $\mathcal{D}_{\text{train+val}}$ using its own hidden-state representations. ARI and NMI measure agreement between the resulting $\mathcal{M}^{(m)}$, while Jaccard matching identifies corresponding clusters and differences in their membership.

As shown in Figure 3, NMI indicates strong agreement between model clusters across all model pairs, with the highest agreement between Qwen3-VL and Qwen3.5. Cross-model NMI consistently exceeds the NMI between model clusters and human cognitive categories reported in Table 1. This pattern extends across model families, suggesting that MLLMs share a cognitive organization that is more similar across models than to the human cognitive taxonomy.

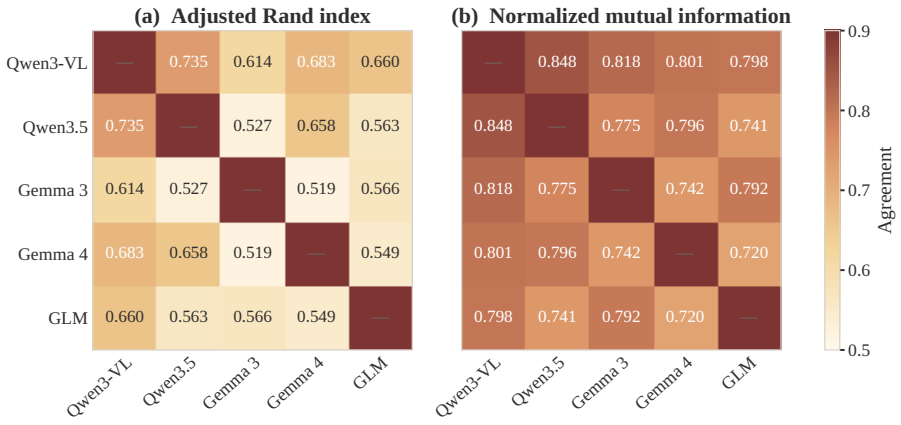



Figure 3: Agreement between model clusters across five MLLMs. Each model forms $K = 24$ clusters from its own representations of the same examples in $\mathcal{D}_{\text{train+val}}$. Panels (a) and (b) report ARI and NMI, respectively. Higher values indicate more similar grouping of the examples.

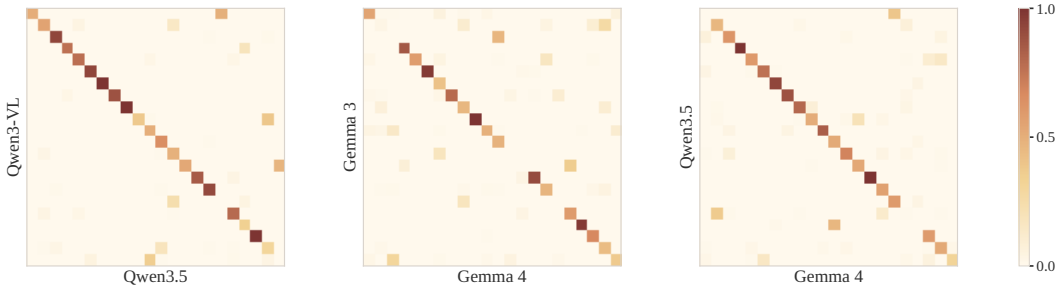



Figure 4: Jaccard overlap between model clusters.

To identify recurring patterns across models, we compare their clusters using Jaccard overlap. We find a one-to-one matching that maximizes the total overlap between the two sets of clusters and arrange matched pairs along the heatmap diagonal. As shown in Figure 4, overlap concentrates along the matched pairs, indicating that different models group many of the same examples together. These results suggest that MLLMs may share a broadly similar cognitive organization while retaining some model-specific boundaries.

To examine which task patterns underlie the cross-model agreement, we inspect examples from clusters with different degrees of membership overlap. As shown in Figure 5 in the appendix, all five models form closely corresponding clusters for object-count comparisons and cube-net folding. Other clusters show weaker cross-model correspondence, including Qwen3-VL clusters involving gear and belt systems and Gemma 4 clusters involving people's intentions. These examples suggest that models recover common groupings for some task patterns, while differing in how they group examples involving other reasoning demands.

### 4.4 Ablation Studies

The ablations examine which choices in cluster construction and mask selection support effective reasoning interventions. We vary cluster granularity, test whether the gains depend on matching masks to clusters, and compare internal and external representations. We also examine whether cluster assignments depend on visual information and whether different extraction layers support effective interventions.

**Number of clusters.** We vary $K$ to examine how cluster granularity affects accuracy and whether model clusters outperform human cognitive categories when their counts match. At $K = 1$, we optimize a single attention-head mask on the entire $\mathcal{D}_{\text{train}}$ and apply it to all examples in $\mathcal{D}_{\text{val}}$ and

$\mathcal{D}_{\text{test}}$. We also evaluate $K \in \{12, 24, 32\}$ for Qwen3-VL, with $K = 12$ matching the twelve human cognitive categories in $\mathcal{H}$.

As shown in Table 5, cluster-specific masks outperform a single shared mask for Qwen3-VL on $\mathcal{D}_{\text{val}}$. With $K = 12$, model clusters also yield higher accuracy than the twelve human cognitive categories. Among $K \in \{12, 24, 32\}$, $K = 24$ achieves the highest test accuracy. Increasing $K$ further does not improve performance. This pattern suggests a trade-off between specialization and generalization. Too few clusters may combine examples with different reasoning patterns, while overly fine clusters may emphasize variation specific to a small set of examples rather than patterns that generalize.

Table 5: Qwen3-VL accuracy (%) using twelve human cognitive categories or $K$ model clusters. Original denotes no intervention. $K = 1$ uses one mask for all examples.

| | Model | Original | Human (12) | $K = 1$ | $K = 12$ | $K = 24$ | $K = 32$ |
|---|---|---|---|---|---|---|---|
| $\mathcal{D}_{\text{val}}$ | Qwen3-VL | 64.16 | 69.38 | 66.73 | 71.69 | **72.56** | 72.05 |
| $\mathcal{D}_{\text{test}}$ | Qwen3-VL | 71.55 | 72.22 | 73.46 | 72.35 | **75.30** | 73.81 |

Table 6: Accuracy (%) on $\mathcal{D}_{\text{val}}$ without intervention (Original), with the mask learned for the example's own cluster, or with masks learned for other clusters. The last column averages over the other $K - 1 = 23$ masks.

| Model | Original | Same-cluster mask | Other-cluster masks |
|---|---|---|---|
| Qwen3-VL | 64.16 | **72.56** | 60.66 |
| GLM-4.6V | 64.14 | **74.80** | 59.33 |

**Cluster-Specific Attention-Head Interventions.** To examine whether reasoning patterns are specific to individual model clusters, we evaluate examples in $\mathcal{D}_{\text{val}}$ using masks learned for the same cluster and masks learned for other clusters. As shown in Table 6, replacing the same-cluster mask with masks from other clusters lowers accuracy below the unmodified model on average for both Qwen3-VL and GLM-4.6V. These results provide functional evidence for cluster-specific reasoning patterns, as head adjustments that support examples within one cluster do not generally benefit examples in other clusters.

**Target Model Hidden States versus External Embeddings.** To examine whether the target model's hidden states better capture its reasoning patterns than external embeddings, we compare Qwen3-VL hidden states with representations from Qwen3-VL-Embedding-2B. Both approaches group the same examples into $K = 24$ clusters. We optimize masks for each set of clusters on Qwen3-VL using the same $\mathcal{D}_{\text{train}}$ and optimization settings. As shown in Table 7, both approaches improve accuracy over the unmodified model on $\mathcal{D}_{\text{val}}$, but interventions guided by Qwen3-VL hidden states achieve higher accuracy. These results suggest that its own hidden states better identify examples with shared reasoning patterns than the external embeddings tested.

Table 7: Qwen3-VL accuracy by representation source on $\mathcal{D}_{\text{val}}$ (%).

| Representation source | Accuracy (%) |
|---|---|
| Qwen3-VL hidden states | **72.56** |
| Qwen3-VL-Embedding-2B | 69.34 |

**Role of Visual Information.** We examine whether examples remain in the same model clusters when their visual information is removed. For each example, we retain the question text, remove the images, and recompute its hidden-state representation. Using the original centroids, we identify the nearest cluster and check whether it differs from the example's cluster with images. As shown in Table 8, many examples move to different clusters across all four tested models. These changes show that visual input contributes to how models represent and organize examples.

**Layers of Hidden States.** We examine whether the layer used to extract hidden states affects how effectively model clusters guide reasoning interventions. We extract hidden states at $\{25\%, 50\%, 75\%, 100\%\}$ of decoder depth, corresponding to decoder layers $\{10, 20, 30, 40\}$ for GLM-4.6V and $\{12, 24, 36, 48\}$ for Gemma 4, respectively. At each extraction layer, we construct $K = 24$ model clusters and optimize a mask for each cluster using the same training procedure.

As shown in Table 9, the highest accuracy occurs at 25% of decoder depth for GLM-4.6V and 50% for Gemma 4, although GLM-4.6V varies little across depths. These results favor earlier or intermediate hidden states for clustering in the evaluated settings. A possible explanation is that information about the question is more accessible in earlier or intermediate hidden states, while later states increasingly focus on next-token prediction. This interpretation is consistent with layerwise attribute-extraction results in language models (Ghandeharioun et al., 2024).

Table 8: Examples that move to a different cluster after image removal (%).

| | Qwen3-VL | Qwen3.5 | Gemma 3 | GLM-4.6V |
|---|---|---|---|---|
| Cluster changes (%) | 100.00 | 100.00 | 64.95 | 64.21 |

Table 9: Accuracy (%) on $\mathcal{D}_{\text{val}}$ using $K = 24$ clusters from hidden states at different decoder depths. Column headings indicate the percentage of decoder depth.

| Model | 25% | 50% | 75% | 100% |
|---|---|---|---|---|
| GLM-4.6V | **74.90** | 74.80 | 74.23 | 74.36 |
| Gemma 4 | 68.78 | **71.15** | 69.21 | 68.88 |

## 5 Conclusion

We examined how human cognitive categories relate to the representational and functional organization of five MLLMs. Model clusters match some human cognitive categories while subdividing others and regrouping examples across categories, with recurring structure across architectures. These representational distinctions also have functional relevance for reasoning, as cluster-guided attention-head interventions outperform those guided by human cognitive categories. They also improve held-out test accuracy over the recent input-dependent steering methods evaluated in our experiments. These findings suggest that human cognitive categories offer a useful but incomplete account of model reasoning, while model clusters reveal additional distinctions that help explain and improve it.

**Limitations.** Our study examines five MLLMs using one developmental cognitive taxonomy and a constructed multiple-choice dataset. Extending the analysis to other tasks and open-ended reasoning remains future work. Representational overlap and intervention effects provide evidence about reasoning patterns, but do not establish independent cognitive mechanisms.

### AI use statement

We used ChatGPT to assist with manuscript polishing. ChatGPT also supported dataset construction, as stated in Appendix A. The authors are responsible for the paper's claims, analyses, and conclusions.

### Ethics statement

This study analyzes pretrained models using existing datasets and synthetic examples. Human cognitive categories serve as comparison references, and our findings do not imply that models possess human cognition. Our planned data release will respect source licenses and redistribution restrictions, with source references for materials that cannot be redistributed.

### Reproducibility statement

Appendix A documents data sources, construction procedures, data splits, and prompts. Appendix B describes the experimental settings, including hidden-state extraction, clustering, mask optimization, and evaluation. We will make the dataset and code publicly available to support reproduction of the experiments and further analysis.

## References

Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Huiyi Chen, Jiawei Peng, Dehai Min, Changchang Sun, Kaijie Chen, Yan Yan, Xu Yang, and Lu Cheng. MVI-bench: A comprehensive benchmark for evaluating robustness to misleading visual inputs in LVLMs. In *International Conference on Machine Learning*, 2026.

Chengqi Duan, Kaiyue Sun, Rongyao Fang, Manyuan Zhang, Yan Feng, Ying Luo, Yufang Liu, Ke Wang, Peng Pei, Xunliang Cai, Hongsheng Li, Yi Ma, and Xihui Liu. CodePlot-CoT: Mathematical visual reasoning by thinking with code-driven images. In *IEEE Conference on Computer Vision and Pattern Recognition Findings*, pp. 9586–9596, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166, 2024.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *International Conference on Machine Learning*, pp. 15466–15490, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Linda Hermer and Elizabeth S Spelke. A geometric process for spatial reorientation in young children. *Nature*, 370(6484):57–59, 1994.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

Brenden M Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International Conference on Learning Representations*, pp. 90960–90985, 2025.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D. Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, and Hokin Deng. Core knowledge deficits in multi-modal language models. In *International Conference on Machine Learning*, pp. 34379–34409, 2025.

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 *European Conference on Computer Vision*, pp. 216–233, 2024.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

Wenxin Ma, Chenlong Wang, Ruisheng Yuan, Hao Chen, Nanru Dai, S. Kevin Zhou, Yijun Yang, Alan Yuille, and Jieneng Chen. CausalSpatial: A benchmark for object-centric causal spatial reasoning. *arXiv preprint arXiv:2601.13304*, 2026.

Xueqi Ma, Jun Wang, Yanbei Jiang, Sarah Erfani, Tongliang Liu, and James Bailey. Cognitive mirrors: Exploring the diverse functional roles of attention heads in LLM reasoning. In *Conference on Neural Information Processing Systems*, pp. 27188–27214, 2025.

Andrew Nam, Henry Conklin, Yukang Yang, Tom Griffiths, Jonathan D Cohen, and Sarah-Jane Leslie. Causal head gating: A framework for interpreting roles of attention heads in transformers. In *Conference on Neural Information Processing Systems*, pp. 115609–115631, 2025.

Jayneel Parekh, Pegah Khayatan, Mustafa Shukor, Arnaud Dapogny, Alasdair Newson, and Matthieu Cord. Learning to steer: Input-dependent steering for multimodal LLMs. In *Advances in Neural Information Processing Systems*, 2025.

Jean Piaget and Bärbel Inhelder. *The Psychology of the Child*. Basic Books, 1969.

Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. CAPTURe: Evaluating spatial reasoning in vision language models via occluded object counting. In *IEEE International Conference on Computer Vision*, pp. 8001–8010, 2025.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, YiFan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? In *Annual Meeting of the Association for Computational Linguistics*, pp. 20023–20070, 2025.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. SAT: Dynamic spatial aptitude training for multimodal language models. In *Conference on Language Modeling*, 2025.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 146–162, 2022.

Taiga Shinozaki, Tomoki Doi, Amane Watahiki, Satoshi Nishida, and Hitomi Yanaka. Do large vision-language models distinguish between the actual and apparent features of illusions? In *Annual Meeting of the Cognitive Science Society*, volume 47, pp. 4339–4345, 2025.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. In *International Conference on Machine Learning*, 2026.

Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles-a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

Haoran Sun, Bingyang Wang, Suyang Yu, Yijiang Li, Qingying Gao, Haiyun Lyu, Lianyu Huang, Zelong Hong, Jiahui Ge, Qianli Ma, et al. Probing perceptual constancy in large vision-language models. *arXiv preprint arXiv:2502.10273*, 2025.

Alvin Wei Ming Tan, Sunny Yu, Bria Long, Wanjing Anya Ma, Tonya Murray, Rebecca D Silverman, Jason D Yeatman, and Michael C Frank. DevBench: A multimodal developmental benchmark for language learning. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, pp. 77445–77467, 2024.

Yunlong Tang, Pinxin Liu, Mingqian Feng, Zhangyun Tan, Rui Mao, Chao Huang, Jing Bi, Yunzhong Xiao, Susan Liang, Hang Hua, et al. MMPerspective: Do MLLMs understand perspective? A comprehensive benchmark for perspective perception, reasoning, and robustness. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, 2024.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Alexander Trott, Caiming Xiong, and Richard Socher. Interpretable counting for visual question answering. In *International Conference on Learning Representations*, 2018.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *International Conference on Machine Learning*, pp. 1073–1080, 2009.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Conference on Neural Information Processing Systems*, pp. 95095–95169, 2024.

Peijie Wang, Zhong-Zhi Li, Fei Yin, Dekang Ran, and Cheng-Lin Liu. MV-MATH: Evaluating multimodal math reasoning in multi-visual contexts. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 19541–19551, 2025a.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for LLMs via dynamic steering vectors. In *International Conference on Learning Representations*, pp. 79334–79351, 2025b.

Zheng Wang, Muchen Li, Renjie Liao, and Yan Leng. IDEEA: training-free input-dependent steering via activation cluster matching. *arXiv preprint arXiv:2609.02089*, 2026.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, pp. 62143–62156, 2025.

Xinbo Wu and Lav R Varshney. Transformer-based causal language models perform clustering. In *Findings of the Association for Computational Linguistics: NAACL*, pp. 5362–5387, 2025.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. LogicVista: Multimodal LLM logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. VisuLogic: A benchmark for evaluating visual reasoning in multi-modal large language models. In *International Conference on Learning Representations*, pp. 25966–26003, 2026.

Wencheng Ye, Xiaoyang Yuan, Yi Bin, Hengyu Jin, Liang Peng, Pengpeng Zeng, and Heng Tao Shen. RISER: Orchestrating latent reasoning skills for adaptive activation steering. In *Findings of the Association for Computational Linguistics*, pp. 4627–4644, 2026.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Yiyang Zhou, Haoqin Tu, Zijun Wang, Zeyu Wang, Niklas Muennighoff, Fan Nie, Chaorui Deng, Shen Yan, Haoqi Fan, Yejin Choi, et al. When visualizing is the first step to reasoning: MIRA, a benchmark for visual chain-of-thought. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 26154–26164, 2026.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded question answering in images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4995–5004, 2016.

# A Dataset Construction

## A.1 Sources and Data Splits

We extend CoreCognition using task prototypes from its twelve human cognitive categories (Li et al., 2025). The resulting dataset contains 3,294 examples. We combine 1,170 original examples with 1,124 extension examples to form $\mathcal{D}_{\text{train+val}}$, and use a separate 1,000-example set as $\mathcal{D}_{\text{test}}$. Each example provides visual input, a question, answer options, a correct answer, and a human cognitive category.

The fixed split assigns 1,377 examples to $\mathcal{D}_{\text{train}}$ and 917 to $\mathcal{D}_{\text{val}}$. We originally generated it for the embedding-based $K = 24$ experiment with seed 42 and a target validation fraction of 0.4 within each cluster. All subsequent comparisons reuse these example identities. Appendix B.1 specifies the data used to fit the cluster centroids. Representational comparisons use $\mathcal{D}_{\text{train+val}}$, and mask optimization uses only $\mathcal{D}_{\text{train}}$. Test examples contribute to neither centroid fitting nor mask optimization. Table 10 gives the category counts.

Table 10: Human cognitive category counts in the fixed training and validation pool and the separate test set. Total sums all three splits.

| Human cognitive category | $\mathcal{D}_{\text{train}}$ | $\mathcal{D}_{\text{val}}$ | $\mathcal{D}_{\text{test}}$ | Total |
|---|---|---|---|---|
| *Boundary* | 101 | 66 | 84 | 251 |
| *Conservation* | 60 | 42 | 76 | 178 |
| *Continuity* | 67 | 40 | 36 | 143 |
| *Hierarchy* | 133 | 83 | 90 | 306 |
| *Intentionality* | 211 | 138 | 143 | 492 |
| *Intuitive Physics* | 48 | 29 | 65 | 142 |
| *Mechanical Reasoning* | 121 | 73 | 24 | 218 |
| *Perceptual Constancy* | 135 | 95 | 101 | 331 |
| *Permanence* | 6 | 8 | 60 | 74 |
| *Perspective* | 166 | 110 | 149 | 425 |
| *Spatiality* | 201 | 145 | 112 | 458 |
| *Tool Using* | 128 | 88 | 60 | 276 |
| Total | 1,377 | 917 | 1,000 | 3,294 |

The separate test set contains examples drawn from or adapted from Visual7W (Zhu et al., 2016), SAT-Perspective (Ray et al., 2025), BLINK (Fu et al., 2024), Perceptual Constancy (Sun et al., 2025), CausalSpatial (Ma et al., 2026), CAPTURe (Pothiraj et al., 2025), CLEVR (Johnson et al., 2017), A-OKVQA (Schwenk et al., 2022), HowMany-QA (Trott et al., 2018), FILM (Shinozaki et al., 2025), GRIP-Benchmark-34, MMVP (Tong et al., 2024), and LogicVista (Xiao et al., 2024). Two additional source labels in the manifest identify the Boundary and Conservation synthetic images created for this study rather than external datasets. The manifest records each example's source, cognitive category, task prototype, question, answer options, correct answer, and image.

The sources used to construct the 1,124 training and validation extensions include Visual7W (Zhu et al., 2016), VisuLogic (Xu et al., 2026), CLEVR (Johnson et al., 2017), IconQA (Lu et al., 2021), HallusionBench (Guan et al., 2024), MIRA (Zhou et al., 2026), LogicVista (Xiao et al., 2024), MV-MATH (Wang et al., 2025a), MATH-Vision (Wang et al., 2024), MechGear, VisualPuzzles (Song et al., 2026), SEED-Bench (Li et al., 2024), MathVista (Lu et al., 2024), the occlusion subset of MVI-Bench (Chen et al., 2026), We-Math (Qiao et al., 2025), Math-VR (Duan et al., 2026), and the English subset of MMBench (Liu et al., 2024).

The extension sources also include mechanical-reasoning practice materials from PrepTerminal, Psychometric Solutions and Innovations, WikiJob, and the ASVAB Testing Program.

## A.2 Construction Procedure

We first summarized the twelve human-defined cognitive categories into several task prototypes, using the category definitions and representative examples from the original benchmark. For each

prototype, we described the visual situation and reasoning operation required to answer the question, as well as its boundary with related categories. We encoded these prototype descriptions and inclusion and exclusion criteria in a reusable GPT-assisted curation skill.

We then used the skill to screen examples from the external datasets listed above at the question level. For each candidate, GPT compared the image and original question–answer pair with the prototype catalogue and proposed at most one cognitive category and prototype. An example advanced to review only if the question instantiated that prototype, answering it required the visual input, and the source answer was supported by visible evidence.

Finally, the authors manually reviewed the proposed additions against their images, questions, answer options, and source answers. We checked whether the proposed prototype matched the reasoning required by the question and whether the image supported the stated answer. Ambiguous or unsupported examples were excluded; only examples that passed this review were added to the dataset.

### A.3 Construction Prompts

The *Conservation* images depict an initial and a final state for questions about number, liquid, length, solid quantity, or area. We generated them with the OpenAI built-in image generator. The following is the complete image-generation prompt for example `aug_conservation_ai_000002`:

> Use case: scientific-educational
> Asset type: conservation reasoning dataset image
> Primary request: On the left show exactly 6 identical brass coins in a compact two-by-three rectangle. On the right show exactly the same 6 identical brass coins in a wide semicircular arc. The objects are individually moved; preserve exactly 6 on each side, with identical identity, size, shape, and color. Only spacing and arrangement change.
> Scene/backdrop: a maple dining table near a bright window.
> Style/medium: photorealistic DSLR educational photography with realistic materials, natural texture, and physically plausible contact shadows.
> Composition/framing: landscape two-state process comparison, perfectly top-down; initial state clearly on the left and final state clearly on the right, separated by a vertical divider and a simple left-to-right arrow.
> Lighting/mood: clean natural teaching demonstration with soft light.
> Constraints: every relevant object fully visible and countable; no overlap or cropping; no written words, letters, numbers, captions, measurements, formulas, logos, or watermark; no decorative clutter.

The *Boundary* images cover object segmentation, foreground-background separation, surface assignment, and edge clarity. They were generated with the same image generator. The following is the complete image-generation prompt for example `aug_boundary_ai_000003`:

> Use case: scientific-educational
> Asset type: one image for a visual multiple-choice boundary-reasoning dataset
> Style/medium: realistic studio photograph with natural material texture
> Primary request: Show exactly 3 distinct smooth gray pebbles touching one another in a compact central cluster. Several edges must touch and at least one pair may overlap slightly, but every individual object must retain a continuous visible outer boundary so the exact number can be recovered by boundary tracing.
> Scene/backdrop: plain neutral surface with generous empty margin and no other objects.
> Subject: exactly 3 similarly sized objects; no fragments and no additional objects.
> Composition/framing: centered close view; all objects fully visible; boundaries large enough to inspect.
> Lighting/mood: soft diffuse lighting with gentle contact shadows that clarify separation.
> Constraints: square image; one coherent scene; the requested boundary relation must be visually unambiguous to a careful viewer; no labels, no letters, no num-

> bers, no arrows, no bounding boxes, no captions, no text, no watermark.
> Avoid: decorative clutter, hands, tools, cropped targets, tiny targets, extra objects, deep shadows that hide boundaries, or any visual element that changes the requested count or relation.

The prompts above show the complete image-generation instructions for individual synthetic examples. Each candidate used its own prompt. *Conservation* prompts varied the objects, initial and final states, and scene, while *Boundary* prompts varied the subject, boundary relation, and rendering style. Each image-generation call returned one PNG image. The question, options, and correct answer remained fixed. The calls did not specify a model version, random seed, or sampling parameters.

**Evaluation prompts.** For multiple-choice evaluation, each cyclic option variant is sent as one user message. Its content contains the RGB image or images first, followed by a text item with the following format; the list of permissible letters and the option lines are generated from the options present in that variant:

```
Answer with only the option letter:  A or B.
Do not include the option text or any explanation.

Question:

Is the ball red?

Choices:

A. Yes
B. No
```

This is an illustrative two-option, one-image example. The instruction uses the available letters, which range from two to six options in the evaluation set. For image-valued answer choices, the option text and choice images are interleaved as separate content items after `Choices:`. We render the message using each backbone's native `processor.apply_chat_template` with `tokenize=False` and `add_generation_prompt=True`. We additionally set `enable_thinking=False` for Qwen3.5, Gemma 4, and GLM-4.6V. For each candidate, we append its answer letter to the rendered generation prefix and score its negative log-likelihood; the lowest-scoring letter is the prediction.

# B Implementation Details

We specify how we construct model representations, optimize attention-head masks, and implement the controls used in the main experiments. All conditions retain the data splits defined in Appendix A.1.

## B.1 Hidden-State Extraction and Clustering

**Representation construction.** We extract representations from each frozen, unmodified model, before any attention-head mask is applied. Each example is rendered with the model's native multimodal generation template. We cyclically rotate the option texts over at most four answer positions, run one prompt-only forward pass for each rotation, and retain the decoder-layer output at the last valid prompt token immediately before answer generation. We average these states across rotations and $\ell_2$-normalize the resulting vector. Neither the gold answer, cognitive category label, teacher rollout, nor a trained mask is used to construct the representation. We use raw hidden states without PCA or another learned dimensionality reduction.

The model-specific extraction layers and representation dimensions are shown in Table 11. Layer indices count decoder layers from one; the denominator gives the total number of language-decoder layers. The last column gives the number of independently optimized mask coefficients in one mask for a human cognitive category or model cluster.

**Clustering.** We fit scikit-learn KMeans separately for each backbone on 2,466 examples, comprising the 2,294 non-video examples in $\mathcal{D}_{\text{train+val}}$ and 172 video examples. Representational comparisons and intervention experiments use only non-video examples. This restriction keeps the input modality consistent across the representational comparisons and intervention experiments. The main setting uses $K = 24$, random seed 42, and the Lloyd algorithm. KMeans minimizes squared Euclidean distance between normalized representations and their centroids. We retain the fitted centroids without renormalizing them and keep the clusters fixed during mask training. Validation inputs contribute to these clusters but not mask optimization. Test inputs contribute to neither.

**Centroid routing.** For a test example, we repeat the same model-specific prompt, rotation, pooling, and normalization procedure, and choose the fixed raw KMeans centroid with minimum squared Euclidean distance. If that cluster has no trained mask because it contains no eligible training examples, routing chooses the nearest centroid among clusters with a saved mask and records this as a fallback. No such fallback occurs in the reported 1,000-example results.

### B.2 Attention-Head Masking

**Mask parameterization.** Following CHG (Nam et al., 2025), we optimize one soft attention-head mask for each human cognitive category or model cluster. As in the main text, $h$ indexes supported attention heads across all language-decoder layers. We parameterize each mask coefficient $g_h$ with a scalar logit $a_h$,

$$g_h = \text{clip}_{[0,1]}[(\zeta - \gamma)\,\sigma(a_h/\beta) + \gamma]\,, \qquad \gamma = -0.1, \quad \zeta = 1.1, \quad \beta = 0.25.$$

Here $\sigma$ is the logistic sigmoid and $\text{clip}_{[0,1]}$ restricts the result to $[0, 1]$. The coefficient scales the head output deterministically. We initialize each logit at the finite upper boundary that gives $g_h = 1$, so the initial mask recovers the unmodified model. After each optimizer update, we project logits above this boundary back to it. Backbone weights remain frozen.

**Answer loss and accuracy.** For an example with $q \in \{2, \ldots, 6\}$ options, we construct $q$ cyclic variants so that the correct answer occupies each answer position once. This differs from representation extraction, which averages at most four rotations. For each variant, we append each candidate answer letter to the model's generation prompt and compute its mean token negative log-likelihood (NLL). If $n_r$ is the NLL for candidate $r$, the answer loss is cross-entropy over scores $-n_r/\tau$, with $\tau = 1$, using the correct candidate as the target.

At evaluation, we select the candidate with the lowest NLL. Accuracy pools all cyclic variants, so an example with $q$ options contributes $q$ predictions. We compute accuracy by scoring candidate answers rather than generating answers freely.

**Optimization and checkpoint selection.** Each mask uses only the examples from its human cognitive category or model cluster in $\mathcal{D}_{\text{train}}$. We run 25 full-data optimization steps with AdamW, learning rate 0.05, zero weight decay, gradient-norm clipping at 1.0, and seed 0. Training uses deterministic algorithms.

From the first step, we add a regularizer with coefficient $\lambda = 2 \times 10^{-4}$. We use a signed logit regularizer, computed as the sum of mask logits after clipping them to the finite active range. Minimizing it encourages lower mask coefficients. The main-text term $R(\mathbf{g})$ denotes this penalty through the logit parameterization.

Table 11: Model-specific representation and mask dimensions.

| Model | Extraction layer | Hidden dim. | Mask params. |
|---|---|---|---|
| Qwen3-VL-8B-Instruct | 18/36 | 4096 | 1152 |
| Gemma 3 12B IT | 24/48 | 3840 | 768 |
| Gemma 4 12B IT | 24/48 | 3840 | 768 |
| Qwen3.5-9B | 16/32 | 4096 | 896 |
| GLM-4.6V-Flash | 20/40 | 4096 | 1280 |

All reported conditions use the final step-25 mask, with this choice fixed in advance rather than selected by validation accuracy. The all-one mask gives the unmodified-model reference.

### B.3 Examples of Shared and Model-Specific Clusters

Figure 5 uses the five fixed $K = 24$ assignments on the same 2,294 non-video examples in $\mathcal{D}_{\text{train+val}}$. The Jaccard scores compare complete cluster memberships; the pictured questions illustrate the selected clusters. Cluster numbers identify within-model assignments and are not comparable across models. The figure abbreviates $M_k^{(m)}$ as C$k$ for each model.

The object-count family has 116 members assigned to the corresponding clusters in all five models; the cube-net family has 154. Their mean pairwise Jaccard overlaps are 0.916 and 0.816, respectively. The candidate model-specific rows contain Qwen3-VL $M_{18}^{(m)}$ (60 examples), Qwen3.5 $M_{12}^{(m)}$ (37), Gemma 3 $M_{1}^{(m)}$ (164), Gemma 4 $M_{10}^{(m)}$ (34), and GLM-4.6V $M_{15}^{(m)}$ (189). Qwen3.5 $M_{12}^{(m)}$ groups shape-assembly questions; its maximum matched Jaccard of 0.529 is lower than the two shared families' mean pairwise overlaps, but higher than the maximum matches of the other four rows.

### B.4 Input-Dependent Steering Baselines

We adapt L2S (Parekh et al., 2025) and IDEEA (Wang et al., 2026) to Qwen3-VL-8B-Instruct and the same multimodal multiple-choice task used for our head masks. Both baselines use the fixed 1,377-example training split for fitting, the 917-example validation split for hyperparameter selection, and the disjoint 1,000-example test set only for final evaluation. All methods use the original multiple-choice prompt and select the candidate answer letter with the lowest NLL at test time. Neither steering baseline uses our cluster assignments or learned masks.

**L2S adaptation.** We adapt L2S (Parekh et al., 2025) to the multiple-choice task by predicting an input-specific steering vector from the unsteered prompt representation. For each cyclic training variant, we construct a target from the candidate-letter scores rather than from generated continuations. Let $s_j(\delta)$ be the negative NLL of candidate letter $j$ when a residual shift $\delta$ is applied at the last prompt token of decoder layer 24, and let $y$ be the correct letter. We use the smooth-margin objective

$$L(\delta) = \tau \log \sum_{j \neq y} \exp(s_j(\delta)/\tau) - s_y(\delta), \qquad \tau = 5,$$

and set the steering target to

$$\mathbf{t} = -57 \, \frac{\nabla_\delta L(0)}{\|\nabla_\delta L(0)\|_2}.$$

A two-layer Tanh network with a bottleneck dimension of 100 predicts this target from the unsteered layer-24 prompt representation. The 1,377 training examples yield 4,470 cyclic variants. We split the training set by example, using 1,102 examples to fit the predictor and 275 to select its checkpoint (epoch 93). At inference, the predicted vector is normalized to an $\ell_2$ norm of 57 and added at the last prompt-token residual position before scoring the candidate letters. The gold answer is used to construct training targets, but is not provided to the predictor at inference. This adaptation does not use free generation.

**IDEEA adaptation.** We fit correctness probes and positive/negative per-head activation clusters from training examples, then match clusters to construct head-level steering directions. At test time, each input selects a direction from its own head activation, without using the candidate or gold answer letter. Validation selects four clusters, 64 heads, and strength 2. The original method targets text-only truthfulness and safety tasks; here positive and negative completions denote correct and incorrect answers to the multimodal multiple-choice prompt. We apply the selected direction at the final prompt-token head position used to score candidate letters.

Both adapted baselines add activation vectors, whereas our masks scale attention-head outputs. Table 4 therefore compares complete intervention methods under the shared backbone, data identities, prompt, and metric, rather than attributing the accuracy difference to any one design choice. We selected every baseline setting on validation data before evaluating the fixed test set.

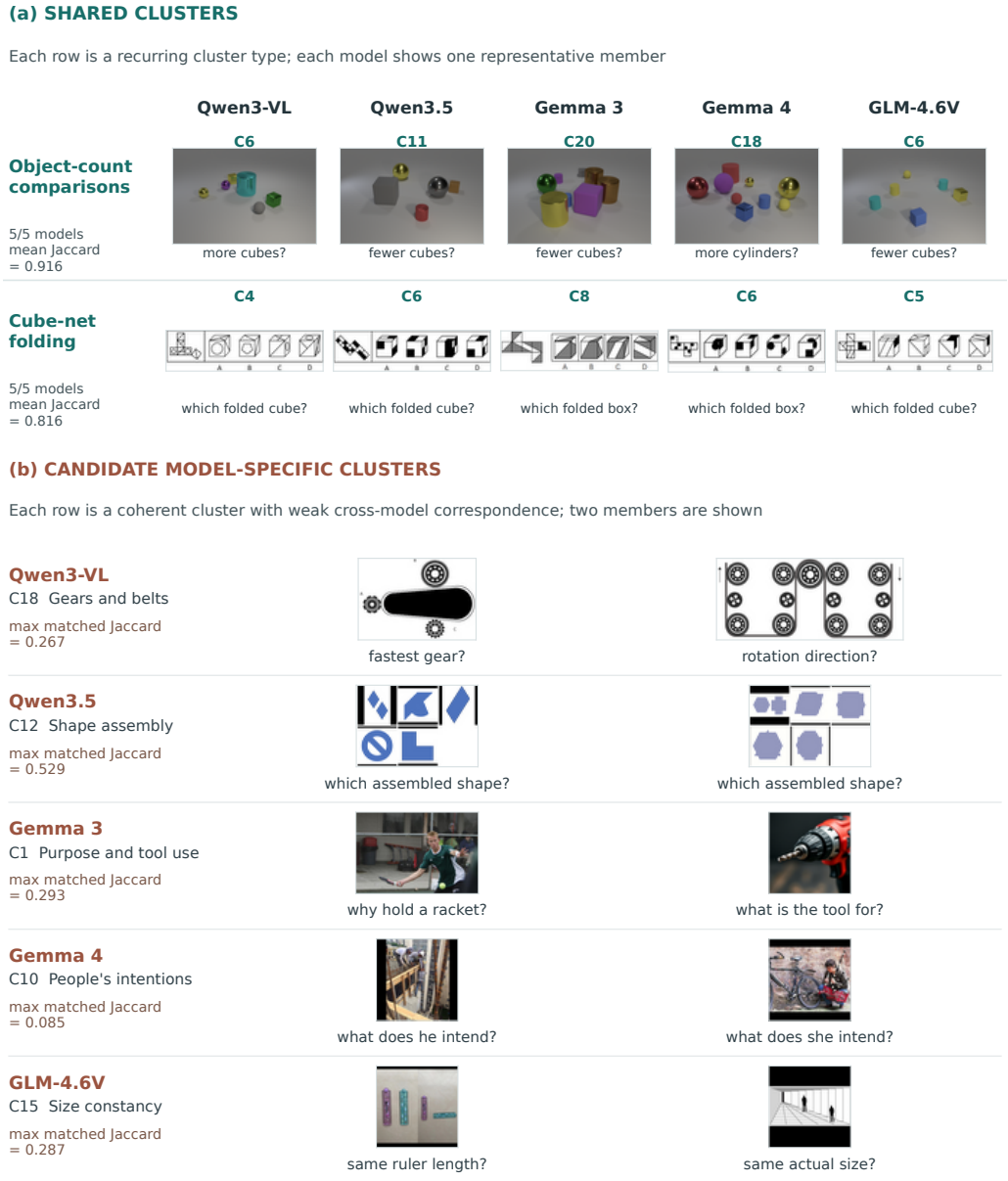



Figure 5: Representative shared and candidate model-specific clusters across MLLMs. (a) Shared clusters. Each row shows a recurring cluster type with strong cross-model correspondence and one representative member of each model's matched cluster. Mean Jaccard averages overlap over the ten matched model pairs. (b) Candidate model-specific clusters. Each row shows two members of a semantically coherent cluster with weaker correspondence to the other models. Max matched Jaccard is the largest, over the four other models, of the best cluster match in each model. All scores use complete cluster memberships over 2,294 aligned examples; the pictures are illustrative. Model-specific refers to grouping boundaries, not a model-exclusive ability or mechanism.

### B.5 Representation and Cluster-Count Controls

**External representation source.** We use the separate model Qwen3-VL-Embedding-2B to represent the same examples and construct an alternative set of $K = 24$ clusters. Qwen3-VL remains the target backbone for mask training and evaluation. This control changes the representation source while retaining the cluster count, target model, data split, and mask optimization. For each example, we serialize the text as `Question:`, the cleaned question, a blank line, `Options:`, and the available lettered options. For image examples, we pass this text and the image path as a text-image input to `SentenceTransformer.encode`; multiple images are combined into one image card. Examples without an image are encoded as text alone. The exact embedding instruction is:

> Represent this visual reasoning question for unsupervised clustering by its visual configuration, question, answer choices, and abstract reasoning required to answer it. Do not infer or encode the correct answer or an option position.

The model's Sentence Transformers configuration uses last-token pooling. We request normalized embeddings from `encode`, explicitly $\ell_2$-normalize the resulting vectors again, and fit Lloyd KMeans in the raw embedding space with $K = 24$ and seed 42. Gold answers, cognitive-category labels, teacher rollouts, and masks are not supplied to the embedding model.

### B.6 Visual Input and Extraction Depth

**Image removal.** We remove images while retaining the text and recompute each representation. For example $i$, let $\mathbf{h}_i^{\text{full}}$ and $\mathbf{h}_i^{\text{text}}$ denote its representations with and without images, and let $\boldsymbol{\mu}_{k_i}$ be its original cluster centroid. We measure the change in cosine distance as

$$\Delta d_i = \left(1 - \cos(\mathbf{h}_i^{\text{text}}, \boldsymbol{\mu}_{k_i})\right) - \left(1 - \cos(\mathbf{h}_i^{\text{full}}, \boldsymbol{\mu}_{k_i})\right).$$

Here cos denotes cosine similarity. Positive distance drift indicates movement away from the original centroid. We also report the percentage of examples that move to a different cluster. We keep the original centroids fixed throughout this analysis.

Table 12 reports the distance drift. Because each model has its own representation geometry, the distance magnitudes do not provide a calibrated measure of visual dependence across models.

**Extraction depth.** We repeat clustering and mask training at decoder layers 10, 20, 30, and 40 for GLM-4.6V-Flash and decoder layers 12, 24, 36, and 48 for Gemma 4 12B IT. These correspond to 25%, 50%, 75%, and 100% of decoder depth. Each condition uses a separate set of $K = 24$ model clusters and separately trained masks. Table 13 compares the GLM-4.6V assignments across depths. The GLM-4.6V accuracy comparison in Table 9 uses validation examples common to all four depth conditions, so its values need not equal the main full-validation results.

Table 12: Mean cosine distance to the original centroid with and without images. Mean drift averages the per-example change $\Delta d_i$.

| Model | Image-Text | Text only | Mean drift |
|---|---|---|---|
| Qwen3-VL | 0.003151 | 0.039470 | +0.036320 |
| Qwen3.5 | 0.016484 | 0.420055 | +0.403571 |
| Gemma 3 | 0.000079 | 0.000215 | +0.000137 |
| GLM-4.6V | 0.021951 | 0.059271 | +0.037321 |

Table 13: Agreement between GLM-4.6V model clusters across relative decoder depths.

| Depth pair | ARI | NMI |
|---|---|---|
| 25%, 50% | 0.6823 | 0.8343 |
| 25%, 75% | 0.6492 | 0.8179 |
| 25%, 100% | 0.6908 | 0.8243 |
| 50%, 75% | 0.7422 | 0.8591 |
| 50%, 100% | 0.6024 | 0.7891 |
| 75%, 100% | 0.6769 | 0.8423 |

# C Additional Representational Analyses

## C.1 Overlap between Human Cognitive Categories and Model Clusters

**Overlap measures.** For a human cognitive category $H_j$ and model cluster $M_k^{(m)}$, we define

$$\text{Jaccard}(H_j, M_k^{(m)}) = \frac{|H_j \cap M_k^{(m)}|}{|H_j \cup M_k^{(m)}|}.$$

Jaccard uses the union as its denominator and gives a symmetric measure of overlap. The mean best Jaccard score in Table 1 averages $\max_k \text{Jaccard}(H_j, M_k^{(m)})$ over the twelve human cognitive categories. These summaries describe membership overlap rather than independent cognitive mechanisms.

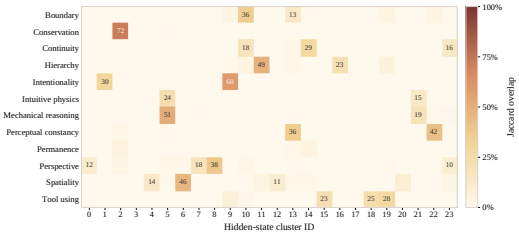



Figure 6: Jaccard overlap between human cognitive categories and Qwen3.5 model clusters. Cell labels show percentages for overlaps of at least 10%.




Figure 7: Jaccard overlap between human cognitive categories and Gemma 3 model clusters. Cell labels show percentages for overlaps of at least 10%.

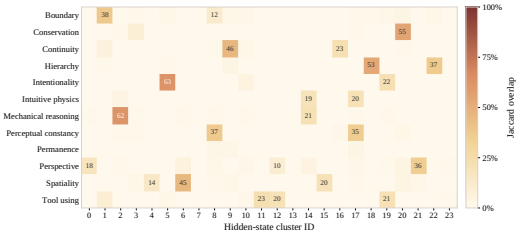



Figure 8: Jaccard overlap between human cognitive categories and Gemma 4 model clusters. Cell labels show percentages for overlaps of at least 10%.

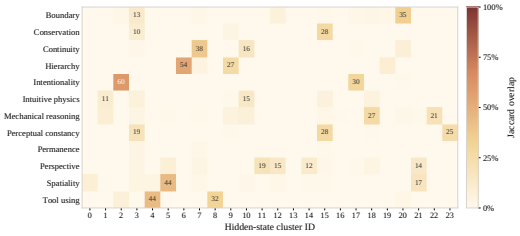



Figure 9: Jaccard overlap between human cognitive categories and GLM-4.6V model clusters. Cell labels show percentages for overlaps of at least 10%.

**Membership visualization and cross-model matching.** For Figure 2, we encode each $H_j$ and $M_k^{(m)}$ as a binary membership vector over the same 2,294 examples. We apply t-SNE to their pairwise Jaccard distances, defined as one minus Jaccard overlap. This visualizes relationships between sets of examples, not distances between their hidden-state representations. For Figure 4, we compute Jaccard overlap between clusters from two models and select a one-to-one matching that maximizes total overlap. Cluster indices have no shared meaning across backbones.

## C.2 Representative Examples

**Representative Qwen3-VL examples.** Figures 10-12 illustrate close alignment, subdivision within a human cognitive category, and regrouping across category boundaries for Qwen3-VL. We use the same $M_k^{(m)}$ indices as in Figures 1 and 2. Each image is paired with a brief summary of its question; the captions report membership counts over $\mathcal{D}_{\text{train+val}}$.

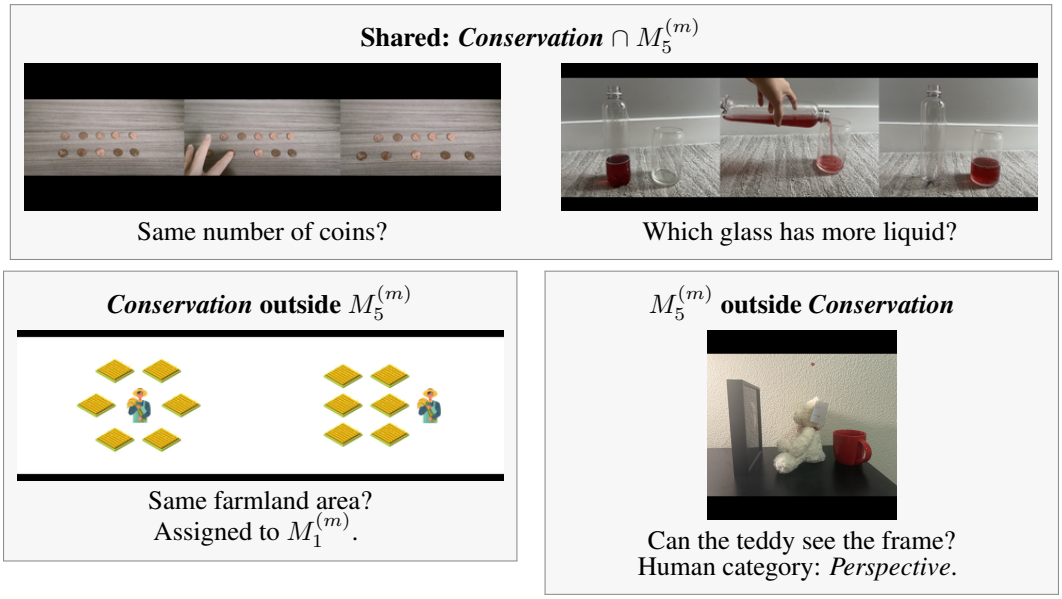



Figure 10: Close but incomplete alignment between *Conservation* and Qwen3-VL cluster $M_5^{(m)}$. They share 95 examples out of 102 in the category and 128 in the cluster, giving Jaccard overlap 0.704. The panels distinguish shared examples from examples belonging to only one of the two sets.

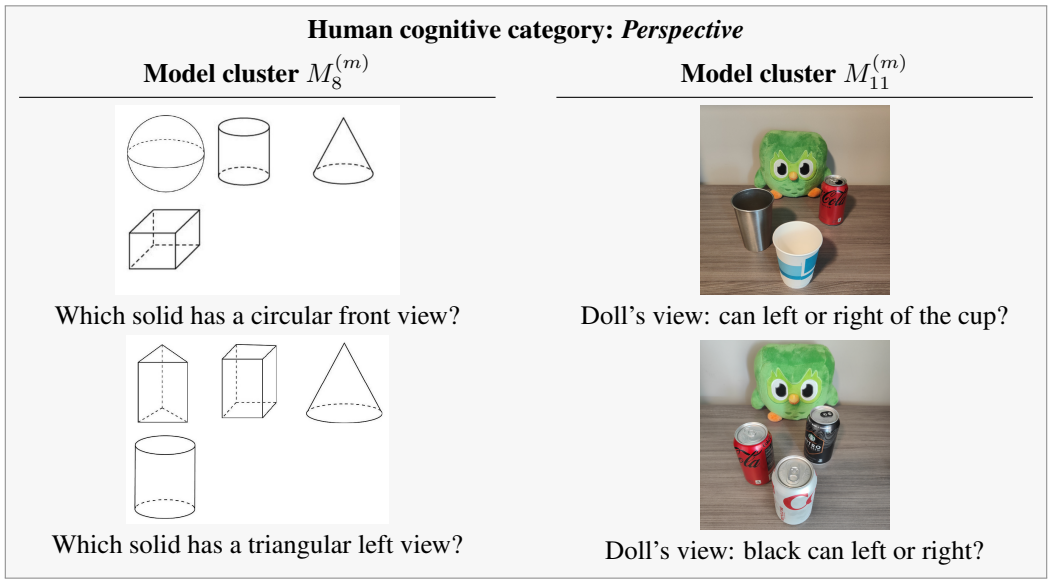



Figure 11: Subdivision within *Perspective*. Cluster $M_8^{(m)}$ contains questions about geometric projection, while $M_{11}^{(m)}$ contains questions about spatial relations across reference frames. They contain 57 and 53 of the 276 *Perspective* examples, respectively.

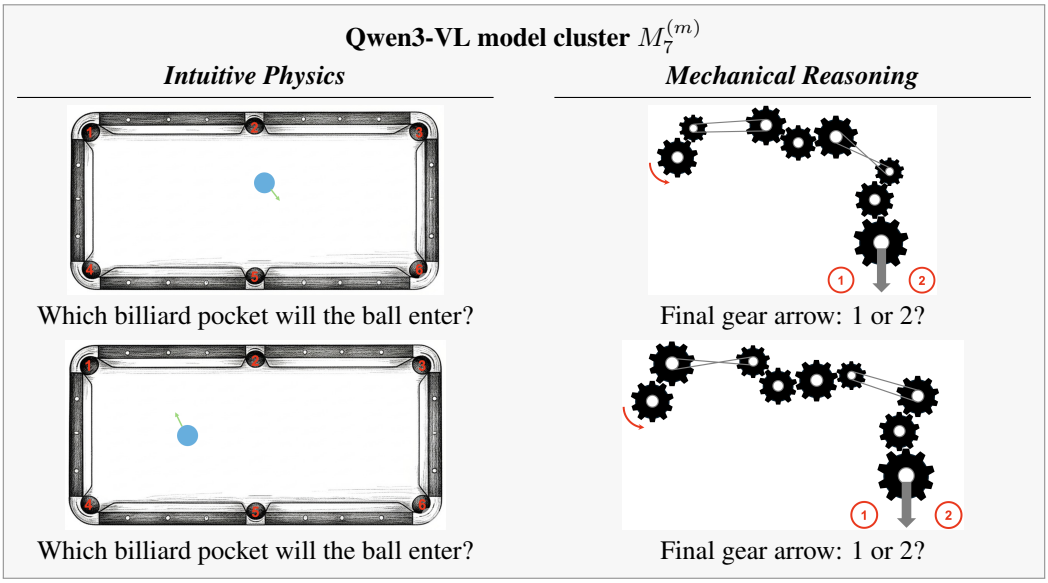



Figure 12: Regrouping across human cognitive categories in Qwen3-VL cluster $M_7^{(m)}$. Its 57 examples comprise 17 *Intuitive Physics* examples and 40 *Mechanical Reasoning* examples. The grouped images show the different questions represented within one model cluster.