Toward Cognitive Supersensing in Multimodal Large Language Models

Boyi Li[1,*] **Yifan Shen**[1,*,§] **Yuanzhe Liu**[1,*] **Yifan Xu**[1] **Jiateng Liu**[1] **Xinzhuo Li**[1] **Zhengyuan Li**[1] **Jingyuan Zhu**[2] **Yunhan Zhong**[1] **Fangzhou Lan**[2] **Jianguo Cao**[2] **James M. Rehg**[1] **Heng Ji**[1] **Ismini Lourentzou**[1,†] **Xu Cao**[1,2,†]

[1]University of Illinois Urbana-Champaign [2]PediaMed AI
[*]Equal contribution [§]Project lead [†]Corresponding author

lourent2@illinois.edu, xucao@pediamed.ai

Abstract

Multimodal Large Language Models (MLLMs) have made rapid progress in open-vocabulary visual understanding, yet they often struggle with cognitive visual reasoning tasks that require maintaining, transforming, and simulating abstract visual states. Existing approaches primarily scale Chain-of-Thought (CoT) reasoning in language, forcing inherently spatial or visual computations into discrete text tokens. We introduce Cognitive Supersensing, a training paradigm that equips MLLMs with latent visual reasoning by integrating a Latent Visual Imagery Prediction (LVIP) head that progressively learns refined visual latent states aligned with the evolving reasoning context and target answer, thereby providing a structured visual grounding signal for internal reasoning. We further introduce a reinforcement learning stage that optimizes rationale trajectories using both answer evidence and LVIP-based latent grounding, encouraging reasoning paths that remain visually consistent. To evaluate the cognitive capabilities of MLLMs, we present CogSense-Bench, a comprehensive visual question answering (VQA) benchmark assessing five cognitive dimensions. Extensive experiments demonstrate that MLLMs trained with Cognitive Supersensing significantly outperform state-of-the-art baselines on CogSense-Bench and exhibit superior generalization on out-of-domain mathematics and science VQA benchmarks, suggesting that internal visual imagery may be key to bridging the gap between perceptual recognition and cognitive understanding. Causal analyses through noise injection, latent patching, and linear probing show that LVIP states progressively encode task-relevant visual structure and influence downstream answer probabilities, supporting their role as active reasoning signals. Project page: https://pediamed.ai/Cognition-MLLM/cogsense/

1 Introduction

Multimodal Large Language Models (MLLMs) have rapidly advanced open-vocabulary visual understanding, enabling strong performance across recognition, grounding, and compositional perception tasks [147]. Yet, existing models can describe what is present but struggle to mentally operate on visual information, *e.g.*, by explaining abstract layouts, simulating future transformations, or inferring visual rules in diagrams [111]. These failures expose a gap between high-level cognitive reasoning and low-level perception; this gap has become a key bottleneck in modern VQA and visual reasoning benchmarks [89, 18] and even extends to reasoning-driven image generation [87, 123].

Current approaches often attempt to narrow this gap by eliciting *Chain-of-Thought* (CoT) reasoning as natural-language reasoning traces, encouraging models to articulate intermediate steps explicitly

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

[139, 176, 124], or by relying on external tools [51]. However, even in multimodal settings, the intermediate computation is externalized in text, which is a poor interface for many visuospatial operations. Many subroutines underlying abstract visual reasoning, such as mentally rotating shapes, simulating dynamics, or inducing rules in pattern matrices, are most naturally expressed as geometric transformations, continuous states, or structured visual relations rather than as a sequence of discrete tokens. Expressing these intermediate states solely in linear text can introduce representational bottlenecks, where spatial relations are compressed into discrete tokens, increasing the risk of information loss and brittle reasoning.

To quantify this cognitive gap, we introduce **CogSense-Bench**, a comprehensive benchmark that operationalizes visual cognition along five core dimensions: fluid intelligence [23, 45], crystallized intelligence [23, 109], visuospatial cognition [140, 11], mental simulation [8], and visual routines [129], which together underpin intuitive theories of physics and psychology [60]. Using CogSense-Bench, we systematically re-evaluate state-of-the-art (SoTA) MLLMs and find that substantial weaknesses persist across all dimensions, even when models are equipped with CoT prompting, suggesting that text-only reasoning is indeed often a brittle interface for tasks that require multi-step visual transformation and manipulation.

We therefore seek to shift part of the intermediate reasoning from discrete tokens to a representation space that better preserves geometry, continuity, and structured visual relations. This direction is consistent with cognitive science accounts of a visuospatial sketchpad (the "mind's eye") that supports maintaining and transforming internal visual representations during problem solving [43, 144, 125], and with recent progress showing that latent visual states can serve as effective substrates for prediction and world modeling [148, 119], as well as with brain-inspired designs that draw on hippocampal mechanisms to enhance video vision language models [16]. Motivated by these insights, we explore whether equipping MLLMs with an internal visual reasoning substrate can better support multi-step, non-linguistic manipulation of visual information. To this end, we introduce **Cognitive Supersensing**, a training paradigm that encourages MLLMs to construct latent image-based internal reasoning chains. Concretely, we augment an MLLM with a **Latent Visual Imagery Prediction (LVIP)** head that predicts progressively refined latent visual representations, i.e. latent "imagery" states that encode intermediate visual reasoning, and aligns them with image representations derived from the target answer and supervision signal, thereby grounding the latent chain toward the correct solution. This design enables multi-step reasoning in a representation space that better preserves visual structure, while still allowing language to provide high-level semantic guidance and final explanations.

To train models under this paradigm, we further curate the train split of CogSense-Bench, a targeted training set spanning five categories and eleven sub-tasks, and adopt a multi-stage training pipeline. First, in a supervised fine-tuning stage, we optimize an **LVIP** objective that provides an auxiliary learning signal encouraging intermediate latent visual states to stay predictive of the final answer's visual representation, tightly coupling the reasoning process with the solution. Second, we introduce a reinforcement learning stage that optimizes rollout trajectories in a latent space conditioned on these visual cognitive latents, encouraging coherent latent dynamics and discouraging brittle text-only reasoning paths. Across extensive experiments, CogSense-8B achieves SoTA performance on all tasks and the strongest overall accuracy (**73.8**), surpassing GPT-5.2 by **+33.5**, while also demonstrating improvements in out-of-domain (OOD) generalization, outperforming strong baselines on challenging science and open-ended VQA benchmarks. Taken together, our results suggest that latent visual imagery can serve as an effective internal scaffold that bridges perceptual recognition and cognitive-level multimodal reasoning. Overall, our contributions are:

- We propose **Cognitive Supersensing**, a latent-space reasoning-and-learning framework with supervised fine-tuning and reinforcement learning stages that endows MLLMs with visual imagery capability by aligning semantic reasoning with latent visual world modeling.
- We introduce **CogSense-Bench**, a comprehensive benchmark spanning five cognitive dimensions: fluid intelligence, crystallized intelligence, visuospatial cognition, mental simulation, and visual routines, providing a systematic testbed for evaluating visual cognition beyond perceptual recognition and supporting future research in this direction.
- Through extensive experiments, we show that reasoning and planning purely in text space is often insufficient for visual cognition: CogSense-8B achieves SoTA performance on CogSense-Bench and exhibits strong OOD generalization on challenging mathematics and science VQA benchmarks, outperforming competitive baselines that rely on text-only planning.

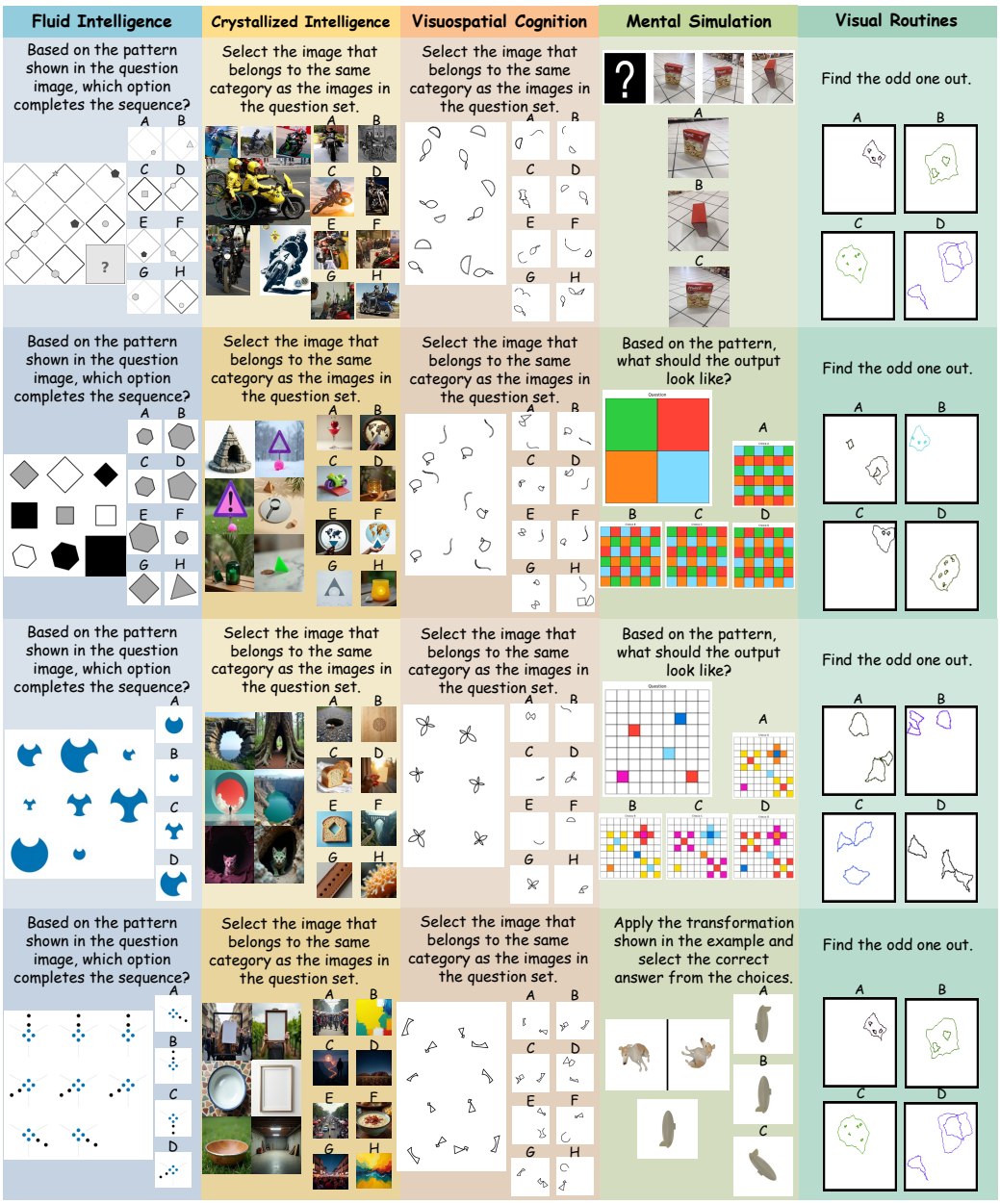

Figure 1: **CogSense-Bench Examples.** Samples from each category of CogSense-Bench. CogSense-Bench comprises various visual cognitive questions classified into five categories: Fluid Intelligence, Crystallized Intelligence, Visuospatial Cognition, Mental Simulation, and Visual Routines, which require visual imagery and cognitive supersensing with deep thinking and reasoning.

2 Related Work

Our work is closely related to Abstract Reasoning [30, 62, 31], Social Reasoning [61, 21, 20, 19], Visual Cognition [12, 13, 108, 107, 106, 7, 163], VLMs for Visual Reasoning [105, 48, 76, 6, 5, 84, 133], and Latent Visual Reasoning [150, 173, 63, 77, 131]. Cognitive Supersensing differs from prior reconstructive and latent-visual-reasoning methods by making latent visual states part of the reasoning trajectory. Instead of reconstructing the input image or using visual latents as generic intermediate scratchpads, LVIP predicts visual hypotheses conditioned on intermediate reasoning prefixes, and the latent-rationale optimization stage uses their consistency to shape the reasoning policy. Additional related work is discussed in Appendix A.

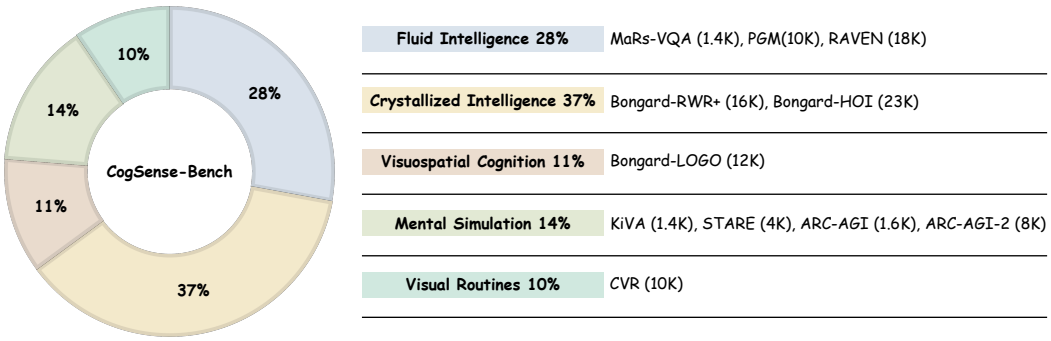

Figure 2: **Data Distribution of CogSense-Bench.** The dataset is structured around five human-cognitive capabilities rather than traditional vision tasks. By aggregating complex abstract reasoning benchmarks, it specifically targets higher-order visual logic and System-2 thinking.

3 CogSense Benchmark

While numerous VQA benchmarks evaluate MLLMs across various tasks, they predominantly focus on semantic recognition or description [79, 29]. There remains a lack of systematic evaluation protocols specifically designed to assess high-level visual cognition [118, 114], i.e. the ability to reason over visual inputs through distinct cognitive mechanisms such as abstract reasoning, spatial structuring, mental simulation, and attention-driven operations. To address this gap, we introduce **CogSense-Bench**, which consists of a comprehensive and large-scale train split together with a unified multi-task test split.

Visual cognition supports high-level visual reasoning and inference processes, such as causal reasoning, intuitive physics, and intuitive psychology [111, 8]. These capacities reflect central dimensions through which humans interpret, predict, and reason about visual environments. To systematically evaluate whether MLLMs exhibit analogous competencies, we assess visual cognitive intelligence across five theory-grounded cognitive categories. These categories are selected to align with established constructs in cognitive science and psychology, and to capture complementary mechanisms underlying human visual reasoning as articulated in prior theoretical and empirical work.

Fluid Intelligence evaluates the capacity to solve unseen reasoning problems independently of prior knowledge [23]. Grounded in Structure Mapping Theory [45], this requires the model to transcend surface attributes and map higher-order logical rules.

Crystallized Intelligence targets the utilization of learned world knowledge [23], relying on Inductive Reasoning and Prototype Theory to abstract semantic concepts from visual variance [109].

To assess 3D spatial understanding, **Visuospatial Cognition** tests the reconstruction of structural relationships, requiring the ability to group discrete visual elements into coherent, holistic structures based on Gestalt laws [140] and the decomposition of shapes into geometric primitives (geons) consistent with Recognition-by-Components Theory [11].

Distinct from static recognition, **Mental Simulation** demands that models act as "simulation engines", employing Hypothetico-Deductive Reasoning to synthesize programs and infer hidden dynamics [8].

Finally, **Visual Routines** evaluates efficiency in visual search, requiring the composition of elementary operations and Focused Attention to bind features and exercise inhibitory control [129].

Figure 1 depicts examples of each category, and Figure 2 shows the data distribution of the train split of CogSense-Bench. The theoretical grounding of the five cognitive dimensions is detailed in Appendix B. Detailed statistics are provided in Appendix C, while Appendix D provides the implementation details for the data pipeline.

CogSense-Bench is designed to probe high-level visual cognition beyond recognition, including abstract rule induction, spatial structuring, mental simulation, and attention control. These categories often require (i) *explicit multi-step reasoning* that composes elementary operations, and (ii) maintaining and manipulating *answer-oriented internal visual states* during inference. Motivated by these requirements, we propose Cognitive Supersensing, which couples chain-of-thought-style rationale generation with latent visual imagery prediction, enabling MLLMs to better support the simulation-like and stateful reasoning demanded by CogSense-Bench.

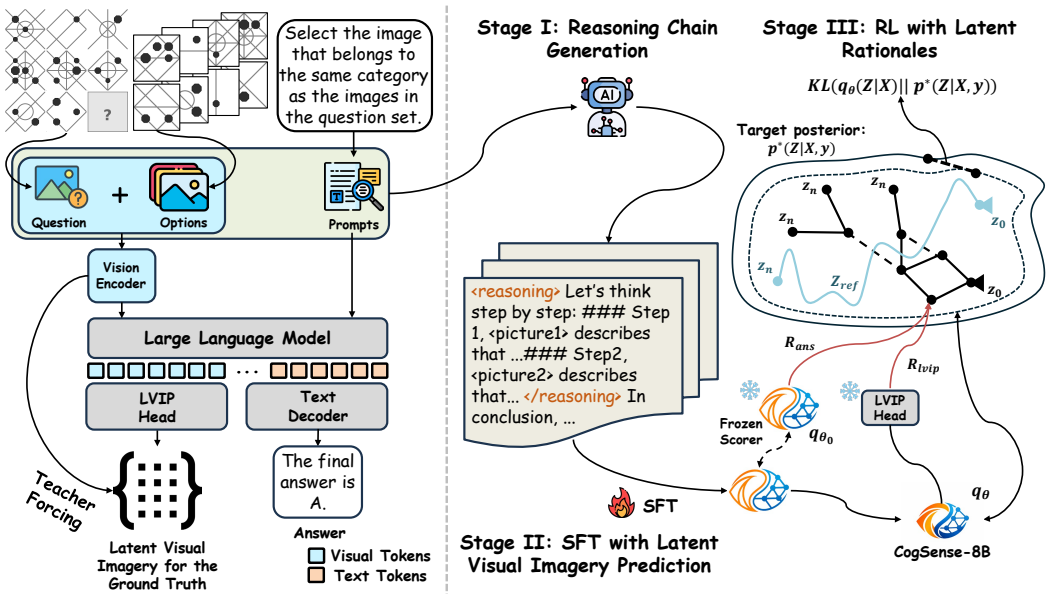

Figure 3: **Cognitive Supersensing framework. Architecture Overview (Left):** CogSense-8B is a VLM that takes images and prompts as input with a text decoder to generate the answer and a Latent Visual Imagery Prediction (LVIP) head to generate a latent visual imagery of the option-image in parallel. **Method Overview (Right):** To train CogSense-8B, we first synthesize teacher-guided reasoning chains (Stage I), then jointly fine-tune the model with text generation and latent visual imagery prediction objectives (Stage II), and finally optimize rationale trajectories with RL using both answer evidence and LVIP-based visual grounding (Stage III).

4 Method

4.1 Preliminaries

Let $\mathcal{V}$ and $\mathcal{Q}$ denote the visual input and the textual prompt, respectively, and denote the multimodal input as $X = (\mathcal{V}, \mathcal{Q})$. Our goal is to learn a multimodal model parameterized by θ, which induces a conditional generation distribution $q_\theta(\cdot \mid X)$ over an output consisting of a reasoning rationale Z (a token sequence) and a final answer y.

Specifically, $\mathcal{V}$ is processed by a pre-trained visual encoder $\mathrm{Enc}_{\mathrm{vis}}(\cdot)$ to extract visual features, denoted as $\mathbf{V}_\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^{T}$ with $\mathbf{v}_i \in \mathbb{R}^{P \times d_{\mathrm{vis}}}$, where T is the number of images, and P is the number of patches per image. These features are mapped into the language embedding space via a projection layer $\mathcal{P}(\cdot)$, yielding projected visual tokens $\mathbf{H}_\mathcal{V} = \mathcal{P}(\mathbf{V}_\mathcal{V})$. Simultaneously, the prompt $\mathcal{Q}$ is tokenized and embedded using the LLM embedding layer to obtain textual tokens $\mathbf{H}_\mathcal{Q}$. The LLM backbone $\mathrm{Enc}_{\mathrm{txt}}(\cdot)$ takes the concatenated token sequence $[\mathbf{H}_\mathcal{V}, \mathbf{H}_\mathcal{Q}]$ as input and produces hidden states used by (i) a text decoder to generate (Z, y) autoregressively and (ii) a Latent Visual Imagery Prediction (LVIP) head to predict answer-oriented latent imagery.

Training follows a three-stage pipeline, shown in Figure 3: **(1)** Reasoning Chain Generation to synthesize high-quality rationales (Section 4.2), **(2)** SFT with Latent Visual Imagery Prediction to jointly learn text generation and latent imagery prediction (Section 4.3), and **(3)** RL with Latent Rationales to further optimize rationale sampling (Section 4.4).

4.2 Stage I: Reasoning Chain Generation

To address the scarcity of high-quality cognitive training data for visual CoT reasoning, we synthesize rationales using a powerful MLLM as a teacher model, denoted as $\mathcal{M}_T$. Given a multimodal input pair $(\mathcal{V}, \mathcal{Q})$, $\mathcal{M}_T$ generates a reasoning rationale Z that demonstrates the logical deduction steps, along with a predicted answer $\hat{y}$, which we then filter to ensure alignment with the ground-truth answer y. Specifically, for each example, we construct a task-specific generation prompt $\mathcal{Q}_{\mathrm{gen}}$. The teacher model is instructed to analyze the visual input and produce a step-by-step reasoning chain based on $(\mathcal{V}, \mathcal{Q}, \mathcal{Q}_{\mathrm{gen}})$:

$$(Z, \hat{y}) \sim \mathcal{M}_T(\cdot \mid \mathcal{V}, \mathcal{Q}, \mathcal{Q}_{\mathrm{gen}}). \tag{1}$$

We filter out generated reasoning chains that fail to reach the correct conclusion ($\hat{y} \neq y$) or contain hallucinated content. This yields an augmented dataset $\mathcal{D}_{\text{chain}} = \{(\mathcal{V}_i, \mathcal{Q}_i, Z_i, y_i)\}_{i=1}^{D}$ for SFT, where D is the total number of examples. Additional details regarding $\mathcal{Q}_{\text{gen}}$ are available in Appendix D.2.

4.3 Stage II: SFT with Latent Visual Imagery Prediction

We propose **Cognitive Supersensing** as an approach for acquiring visual cognitive capabilities. Motivated by the human *constructive matching* process, we introduce an auxiliary module, Latent Visual Imagery Prediction (LVIP), to predict $\mathbf{h}_y$, the latent representation of the ground-truth answer option image. The latent visual imagery generation is synchronized with the autoregressive next-token prediction. At each decoding step of the textual CoT, the LVIP head generates a corresponding visual latent embedding based on the current context.

LVIP Head. The LVIP head $g_\psi(\cdot)$ is a two-layer MLP attached to the shared LLM backbone, in addition to the text decoder. LVIP predicts the latent representation of the answer option using the backbone hidden states corresponding to visual tokens. Let $\mathbf{H}_{\mathcal{V}}^{\text{out}} \in \mathbb{R}^{N \times d}$ be the sequence of hidden states for visual tokens output by $\text{Enc}_{\text{txt}}(\cdot)$, where N is the number of visual tokens and d is the hidden size of the LLM. We extract the subset corresponding to the option images, denoted as $\mathbf{H}_{\text{opt}} \in \mathbb{R}^{M \times d}$, where M is the number of tokens corresponding to option images. We apply average pooling over $\mathbf{H}_{\text{opt}}$ to obtain an aggregated representation $\bar{\mathbf{H}}_{\text{opt}} \in \mathbb{R}^{1 \times d}$, and compute the predicted latent imagery $\hat{\mathbf{h}}_y = g_\psi(\bar{\mathbf{H}}_{\text{opt}}) \in \mathbb{R}^{1 \times d_{\text{vis}}}$.

Learning Objectives. Let v_y denote the candidate option image indexed by the ground-truth answer y. The supervision target for LVIP is the embedding of v_y extracted by the (frozen) visual encoder and then projected into the same space as $\hat{\mathbf{h}}_y$ via an MLP, where $\mathbf{h}_y = \text{MLP}(\mathbf{v}_y) \in \mathbb{R}^{1 \times d_{\text{vis}}}$. We optimize LVIP with an MSE loss between $\hat{\mathbf{h}}_y$ and $\mathbf{h}_y$, jointly with the standard autoregressive cross-entropy loss over the target text sequence $x = (Z, y)$ conditioned on $X = (\mathcal{V}, \mathcal{Q})$:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{|x|} \log q_\theta(x_t \mid X, x_{<t}) + \beta \cdot \text{MSE}(\hat{\mathbf{h}}_y, \mathbf{h}_y), \tag{2}$$

where β balances the two objectives. During inference, the final answer is generated by the text decoder, while the LVIP head is kept as a frozen auxiliary module to produce $\hat{\mathbf{h}}_y$ as a latent visual grounding signal that supports answer-oriented reasoning.

4.4 Stage III: RL with Latent Rationales

In this stage, we leverage diversity-seeking RL by performing amortized variational inference over latent rationales using a Generative Flow Network (GFlowNet) [9, 59]. Concretely, we refine the rationale policy $q_\theta(Z \mid X)$ to sample rationale trajectories Z in proportion to an unnormalized trajectory score, rather than committing to a single deterministic chain. We define the corresponding reward-induced target posterior over rationales as

$$p^*(Z \mid X, y) \;\propto\; \exp(R(Z; X, y)), \tag{3}$$

and train $q_\theta(Z \mid X)$ to approximate this posterior with a GFlowNet objective, detailed below. We define the trajectory score as a weighted combination of answer evidence and LVIP-based representation grounding:

$$R(Z; X, y) = \alpha\, R_{\text{ans}}(Z; X, y) + \gamma\, R_{\text{lvip}}(Z; X, y). \tag{4}$$

Importantly, the LVIP term provides answer-oriented grounding in representation space, which complements discrete answer supervision and facilitates exploration over long rationale trajectories.

Token-Wise Marginal Reward Estimation. Let $Z = (z_1, \ldots, z_L, \top)$ be an autoregressive rationale, with $\top$ denoting end of sequence. We define the answer-evidence component using a frozen scorer to avoid a moving-target reward:

$$R_{\text{ans}}(Z; X, y) = \log q_{\theta_0}(y \mid X, Z), \tag{5}$$

where q_{θ_0} is a frozen scorer (a fixed copy of the supervised model). To incorporate LVIP grounding during RL while keeping it stationary, we keep the LVIP head g_ψ frozen and compute a representation-level reward that depends on Z through the backbone conditioning on $[X; Z]$.

Table 1: **CogSense-Bench Cognitive Ability Results.** CogSense-8B achieves the best results in all five dimensions and the highest Avg. (accuracy over all 1,000 test items), outperforming both closed-source MLLMs and fine-tuned (FT) open-source baselines. Best and second-best results in **bold** and underline.

Model	Fluid Intel.	Crystallized Intel.	Visuospatial Cog.	Mental Simu.	Visual Rout.	Avg.
Human	82.7	91.3	88.5	97.9	78.7	88.4
Gemini 2.5 Flash	23.2	40.2	31.0	40.0	45.2	34.9
GPT-o3	4.7	51.4	20.4	38.7	43.0	32.3
GPT-5.2	29.3	35.9	57.5	60.0	37.6	40.3
Claude Sonnet 4	22.5	31.3	26.5	58.0	34.4	32.6
Grok 4 Fast	13.0	45.4	41.6	21.3	37.6	31.7
Llama-4-Scout-17B	20.3	29.9	35.4	48.7	41.9	31.8
Gemma-3-27B	18.5	29.3	39.8	55.3	43.0	32.7
Ministral-3-14B	9.4	10.6	10.6	39.3	31.2	16.5
Intern-S1	6.5	20.9	12.4	17.3	43.0	17.5
Qwen3-VL-30B	30.8	34.0	37.2	56.0	40.9	37.4
Qwen3-VL-8B (FT)	51.1	76.6	63.7	59.3	41.9	62.3
InternVL3.5-8B (FT)	52.2	80.2	63.7	59.3	41.9	63.9
Qwen3.5-9B (FT)	56.5	89.1	61.9	62.7	46.2	69.1
CogSense-8B (Ours)	**63.8**$_{\uparrow 7.3}$	**91.0**$_{\uparrow 1.9}$	**69.0**$_{\uparrow 7.1}$	**68.0**$_{\uparrow 5.3}$	**50.5**$_{\uparrow 4.3}$	**73.8**$_{\uparrow 4.7}$

Let v_y be the option image indexed by the ground-truth answer y, and let $\mathbf{h}_y = \mathrm{MLP}(\mathbf{v}_y)$ be its embedding from a frozen visual encoder followed by an MLP. Let $\bar{\mathbf{H}}_{\mathrm{opt}}(X, Z; \theta)$ denote the average-pooled final-layer hidden state of the option-image visual tokens produced by the backbone when conditioned on $[X; Z]$. We define:

$$R_{\mathrm{lvip}}(Z; X, y) = -\left\| g_\psi\big(\bar{\mathbf{H}}_{\mathrm{opt}}(X, Z; \theta)\big) - \mathbf{h}_y \right\|_2^2. \tag{6}$$

To provide prefix-level training signals required by flow-based objectives, we define the prefix state at step t as $\tau_t = (z_{1:t}, \top)$, where $\top$ indicates the availability of a termination action rather than an emitted token. Instead of evaluating the scorer at every token, we compute $\widetilde{R}(\tau_t)$ only at sparse *anchor* indices with stride λ and linearly interpolate within each segment. Specifically, let anchors be $t \in \{0, \lambda, 2\lambda, \ldots\}$ with $t < L$, and define $t^+ = \min(t+\lambda, L)$. For any integer $i \in \{0, 1, \ldots, t^+ - t\}$,

$$\widetilde{R}(\tau_{t+i}) = \widetilde{R}(\tau_t) + \frac{i}{t^+ - t}\Big(\widetilde{R}(\tau_{t^+}) - \widetilde{R}(\tau_t)\Big). \tag{7}$$

At anchor positions, we set $\widetilde{R}(\tau_t) = R(\tau_t; X, y)$ with

$$R(\tau_t; X, y) = \alpha \, \log q_{\theta_0}(y \mid X, z_{1:t}) + \gamma \, R_{\mathrm{lvip}}(\tau_t; X, y), \tag{8}$$

where $R_{\mathrm{lvip}}(\tau_t; X, y)$ is computed analogously by conditioning the backbone on $[X; z_{1:t}]$. We employ a standard SubTB loss [90] to train $q_\theta(Z \mid X)$ toward the reward-induced target distribution.

Reference-Guided GFlowNet Fine-tuning. To reduce variance from low-quality samples, we anchor exploration with a reference rationale Z_{ref}. For each X, we sample m candidates $\{Z_i\}_{i=1}^m \sim q_\theta(\cdot \mid X)$ and keep only those that meet a relative evidence threshold:

$$\mathbb{I}(Z_i) = \mathbf{1}[\, R(Z_i; X, y) \;\geq\; R(Z_{\mathrm{ref}}; X, y) + \log \delta_s \,], \tag{9}$$

where $\delta_s \in (0, 1]$ and s is the index of the current training step (so $\log \delta_s \leq 0$ controls the allowed slack in log-space relative to the reference). We optimize SubTB only on accepted trajectories:

$$\mathcal{L}(\theta) = \sum_{i=1}^{m} \mathbb{I}(Z_i) \cdot \mathcal{L}_{\mathrm{SubTB}}(Z_i; \theta), \tag{10}$$

where $\mathcal{L}_{\mathrm{SubTB}}$ is instantiated with the densified prefix scores $\widetilde{R}$.

Bayesian Posterior over Latent Rationales. At inference time, we treat Z as latent and aggregate evidence across multiple sampled rationales. We sample n rationales $\{Z_i\}_{i=1}^n \sim q_\theta(Z \mid X)$, decode an answer y_i conditioned on (X, Z_i), and compute a length-normalized evidence score with the frozen scorer:

$$S_i = \frac{1}{|Z_i| + |y_i|} \log q_{\theta_0}(y_i \mid X, Z_i). \tag{11}$$

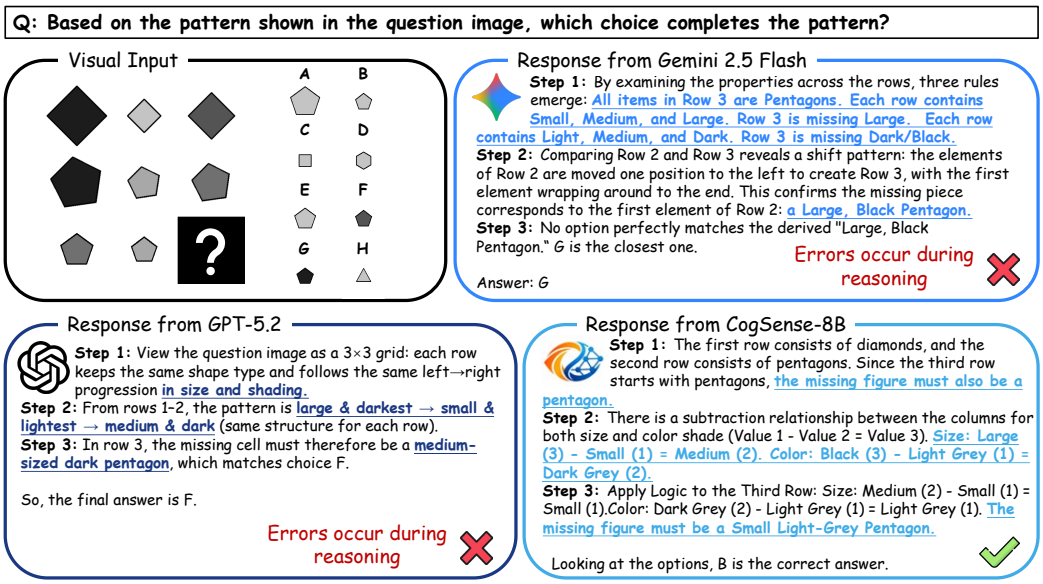

Figure 4: **Qualitative Example of Visual Cognition Reasoning Across Models.** We underline decisive sentences in the reasoning chain. CogSense-8B demonstrates a coherent, multi-step logical chain that closely matches the ground truth, while other models exhibit less precise or less interpretable reasoning paths.

We then output $\hat{y} = y_{i^*}$, where $i^* = \arg\max_i S_i$, which serves as a simple MAP-style selection over sampled latent rationales and reduces sensitivity to any single brittle chain under reasoning ambiguity.

5 Experiments

Baselines and Metrics. We compare CogSense-8B with several mainstream open-source and closed-source MLLMs and VLMs with strong multimodal performance: Gemini 2.5 Flash [32], GPT-o3 [102], GPT-5.2 [101], Claude Sonnet 4 [2], Grok 4 Fast [143], Llama-4-Scout-17B [97], Gemma-3-27B [44], Ministral-3-14B [98], Intern-S1 [4], and Qwen3-VL-30B [5]. To provide a fair comparison, we also fine-tune Qwen3-VL-8B [5], InternVL3.5-8B [134], and Qwen3.5-9B [104] on our train split with the same settings and then evaluate them. We use accuracy as the evaluation metric. We further recruited 20 participants to answer 100 questions selected from CogSense-Bench via stratified sampling, and report their accuracy in Table 1 as a human baseline. Detailed information about the human study can be found in Appendix E. We also compare CogSense-8B with ROSS [131], which reconstructs question images for self-supervised learning. A detailed comparison can be found in Appendix F.

Implementation Details. All experiments are conducted with 8 NVIDIA H200 GPUs. We use Qwen3-VL-8B as the backbone to train CogSense-8B. We use AdamW as the optimizer with a learning rate of $\eta = 10^{-5}$ and a weight decay of 10^{-5}. We use GPT-5 [120] as the teacher model in Stage I.

5.1 Cognitive Ability Results

Quantitative Results. As Table 1 illustrates, CogSense-8B narrows the performance gap relative to existing MLLMs, suggesting that it effectively uses visual imagery to support cognitive reasoning. Despite being an 8B model, it achieves SoTA performance on all tasks and delivers the strongest overall accuracy (**73.8**), surpassing GPT-5.2 by **+33.5**. It also outperforms all fine-tuned baselines, confirming that our strong performance stems from our methodological contributions rather than mere dataset exposure. Notably, for highly symbolic tasks like Visual Routines, which are image-only and are difficult even for humans to solve, the model still outperforms the baselines, indicating that the average pooling operation in LVIP does not discard fine-grained visual information; rather, LVIP acts as a guiding signal. A statistical analysis of the experimental results is provided in Appendix G.

Qualitative Comparison. Figure 4 shows a qualitative comparison between CogSense-8B and mainstream models, demonstrating that Gemini 2.5 Flash and GPT-5.2 both fail to extract the correct underlying rule of the pattern, leading to incorrect answers. In contrast, CogSense-8B demonstrates

Table 2: **General Ability Results.** Performance comparison in general tasks with CogSense-8B and the base model. CogSense-8B demonstrates similar general ability performance to the base model.

Model	GQA	ScienceQA	RealWorldQA	BLINK	HallusionBench	MMSI-Bench	ViewSpatial-Bench	ERQA
Qwen3-VL-8B (Base)	71.4	**92.6**	71.5	64.7	**61.1**	28.0	38.6	46.8
CogSense-8B (Ours)	**71.8**	**92.6**	**71.9**	**65.3**	60.5	**32.4**	**40.8**	**48.8**

Table 3: **Module Ablation Study Results.** We compare the base model and the following variants: SFT *w/o* LVIP, SFT *w/* LVIP, SFT *w/o* LVIP + GRPO, SFT *w/* LVIP + GRPO, SFT *w/o* LVIP + GRPO + Latent Rationales (LR), SFT *w/* LVIP + GRPO + LR, SFT *w/o* LVIP + LR, and CogSense-8B. We **bold** the best results and underline the runner-ups.

Variant	Fluid Intel.	Crystallized Intel.	Visuospatial Cog.	Mental Simu.	Visual Rout.	Avg.
Qwen3-VL-8B (Base)	31.2	34.8	31.0	45.3	40.9	35.5
Qwen3-VL-8B SFT *w/o* LVIP	51.1	76.6	63.7	59.3	41.9	62.3
Qwen3-VL-8B SFT *w/* LVIP	55.4	88.6	61.1	61.3	44.1	68.1
Qwen3-VL-8B SFT *w/o* LVIP + GRPO	55.8	79.9	63.7	63.3	43.0	65.5
Qwen3-VL-8B SFT *w/* LVIP + GRPO	59.1	89.9	64.6	65.3	46.2	70.8
Qwen3-VL-8B SFT *w/o* LVIP + GRPO + LR	56.5	80.4	64.6	64.0	44.1	66.2
Qwen3-VL-8B SFT *w/* LVIP + GRPO + LR	<u>59.8</u>	<u>90.2</u>	<u>65.5</u>	<u>66.7</u>	<u>47.3</u>	<u>71.5</u>
Qwen3-VL-8B SFT *w/o* LVIP + LR	57.2	80.7	65.5	65.3	45.2	66.9
CogSense-8B (Ours)	**63.8**	**91.0**	**69.0**	**68.0**	**50.5**	**73.8**

high-quality reasoning with a concise and clear expression of the underlying pattern. This comparison suggests that CogSense-8B more effectively captures the underlying visual regularities required for abstract pattern reasoning. Additional qualitative examples can be found in Appendix H.

5.2 General Ability Results

We also evaluate the general vision-language understanding and spatial reasoning abilities of CogSense-8B. As shown in Table 2, we evaluate on vision-language datasets such as GQA [53], ScienceQA [88], RealWorldQA [142], BLINK [42], and HallusionBench [46], and spatial-related benchmarks such as MMSI-Bench [149], ViewSpatial-Bench [75], and ERQA [126]. CogSense-8B maintains robust performance on these general benchmarks, comparable to the base model. This balance suggests that Cognitive Supersensing's visual cognitive gains do not stem from overfitting on the training data; instead, our method injects high-level visual cognitive knowledge while preserving the foundational knowledge and instruction-following capabilities inherent in the pre-trained backbone. We attribute the performance decline on the **OCR-related** benchmark, HallusionBench, to a natural post-training shift in which the model trades some specialized text-extraction proficiency for stronger pure visual-spatial logic. Additional general-ability evaluations are in Appendix I.

5.3 Ablation Studies

Module Ablation. As summarized in Table 3, to isolate the contributions of LVIP and our proposed reinforcement learning mechanism, we perform a hierarchical evaluation across various variants. Compared to the vanilla Qwen3-VL-8B, applying standard SFT substantially improves the average accuracy from 35.5 to 62.3. The integration of LVIP introduces a further improvement, raising the average accuracy to **68.0**. This suggests that LVIP helps the model align semantic reasoning and abstract world modeling, enabling the model to gain initial "supersensing" capabilities. To validate the necessity of our specific RL design, we introduce GRPO as a reinforcement learning baseline. Although applying GRPO to the SFT models (*w/o* or *w/* LVIP) yields observable gains, achieving improvements of **+3.2** and **+2.8** compared to models without GRPO, it serves primarily to refine the policy within a standard optimization scope. GRPO plus Latent Rationales further improves performance to **66.2** and **71.5** compared to GRPO only, suggesting that the two RL objectives are complementary. Finally, CogSense-8B achieves the best average performance of **73.8**, outperforming the strongest baseline (SFT *w/* LVIP + GRPO + Latent Rationales) by a clear margin. This result highlights that our elaborately designed RL method is not merely an optimization trick but a specialized reasoning refiner that is more effective than general GRPO in leveraging the visual supersensing established by LVIP.

Latent Visual Imagery Learning. To verify that LVIP learns meaningful intermediate visual states rather than answer-conditioned shortcuts, we perform three experiments. **(i) Noise Injection [81]:** Perturbing the latent state at Step 30 increases the MSE loss between intermediate latent visual imagery at multiple steps of the CoT and the final ground-truth latent visual imagery (Table 4a), demonstrating

Table 4: **Quantitative Analysis of Latent Visual Imagery Learning.** The three sub-tables report the MSE loss after noise injection, probability changes from latent patching, and linear probing accuracy, respectively.

(a) **Noise Injection.**

Intervention	Step 0	Step 10	Step 20	Step 30	Step 40
Add Gaussian Noise @ Step30	0.492	0.368	0.215	0.374	0.418
CogSense-8B *w/o* Noise	0.492	0.368	0.215	0.184	0.158

(b) **Latent Patching.**

Patch Type	Δp
Correct→Incorrect Patch	-0.43
Incorrect→Correct Patch	$+0.22$

(c) **Linear Probing.**

	Step 0	Step 10	Step 20	Step 30	Step 40
Acc.	0.173	0.237	0.394	0.552	0.621

that the uncorrupted intermediate state is vital for reasoning. **(ii) Latent Patching [82]:** Swapping latent states between correct and incorrect runs shifts the rationale and answer probability toward the patched state (Table 4b), showing that these states carry causal information that steers inference. **(iii) Linear Probing [1]:** A linear classifier trained on intermediate latent visual imagery at multiple steps of the CoT shows that accuracy grows progressively (Table 4c). This confirms that visual logic is constructed step-by-step, ruling out trivial answer-conditioned shortcuts. Detailed settings, more experiments, and discussions of latent visual imagery learning can be found in Appendix J.

Additional Ablation Studies. We also conduct ablation studies on CoT (Appendix J), prefix sensitivity (Appendix J), the number of sampled rationales (Appendix K), sampling strategies (Appendix L), teacher models (Appendix M), and backbone models (Appendix N).

5.4 Out-of-Domain Evaluation

To further verify the generalization capability of CogSense-8B, we extend our evaluation to out-of-domain scenarios using the Chemistry and Math subsets of the EMMA benchmark [49], where both the questions and the options contain images. Selected EMMA examples are shown in Figure 5. As shown in Table 5, CogSense-8B achieves substantial gains of **+6.2** and **+8.8** on the Chemistry and Math subsets, respectively, confirming that our method learns generalized visual cognition patterns rather than overfitting to specific training data. We also evaluate CogSense-8B on OOD open-ended VQA benchmarks in Appendix O.

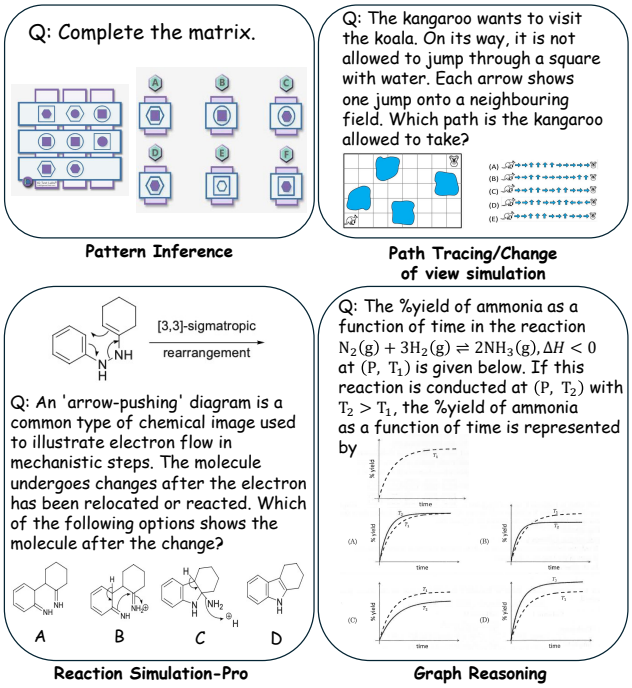

Figure 5: **EMMA Benchmark Sample Problems.**

Table 5: **Out-of-Domain Evaluation**. Generalization ability of CogSense-8B against the backbone on the EMMA benchmark.

Model	Chemistry	Mathematics
Qwen3-VL-8B (base)	39.2	26.0
CogSense-8B (Ours)	**45.4**$_{\uparrow 6.2}$	**34.8**$_{\uparrow 8.8}$

6 Conclusion

We propose Cognitive Supersensing, a paradigm designed to bridge the gap between perceptual processing and complex cognitive reasoning in MLLMs. By integrating a Latent Visual Imagery Prediction (LVIP) head and employing a training strategy that combines SFT and RL with Latent Rationales, we enable CogSense-8B to simulate internal visual imagery aligned with semantic reasoning chains. We further introduce CogSense-Bench, a comprehensive evaluation suite targeting five dimensions of visual cognition. Our experiments demonstrate that CogSense-8B achieves superior performance compared to SoTA MLLM baselines on these cognitive tasks. These results indicate that modeling the interaction between visual simulation and logical deduction is an effective direction for advancing the reasoning abilities of multimodal systems.

Acknowledgments and Disclosure of Funding

We acknowledge the support of the Health Care Engineering Systems Center within the Coordinated Science Lab at UIUC. This research was partially supported by the NSF CAREER Award #2542328, the U.S. Defense Advanced Research Projects Agency (DARPA) under award HR001125C0303, and the U.S. Army under contract W5170125CA160. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, DARPA, the U.S. Army, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. Portions of this work were also supported by Google PhD Fellowship, PediaMed AI and NemX Labs' innovation fund for computational resource and travel.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[2] Anthropic. System card: Claude Opus 4 & Claude Sonnet 4. System card (technical report), May 2025. URL `https://www.anthropic.com/claude-4-system-card`. Document shows changelog updates on July 16, 2025 and September 2, 2025.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[4] Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, et al. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*, 2025.

[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. *arXiv preprint arXiv:2511.21631*, 2025.

[6] 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. *arXiv preprint arXiv:2502.13923*, 2025.

[7] David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning*, pages 511–520. PMLR, 2018.

[8] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45): 18327–18332, 2013.

[9] Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. Gflownet foundations. *Journal of Machine Learning Research*, 24(210):1–55, 2023.

[10] Yaniv Benny, Niv Pekar, and Lior Wolf. Scale-localized abstract reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12557–12565, 2021.

[11] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987.

[12] Mikhail Moiseevich Bongard. The recognition problem. Technical report, Foreign Technology Div Wright-Patterson AFB Ohio, 1968.

[13] Mikhail Moiseevich Bongard. *Pattern Recognition*. Spartan Books, 1970. ISBN 978-0-87671-118-7.

[14] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.

[15] Giacomo Camposampiero, Michael Hersche, Roger Wattenhofer, Abu Sebastian, and Abbas Rahimi. I-RAVEN-X: Benchmarking generalization and robustness of analogical and mathematical reasoning in large language and reasoning models. *arXiv preprint arXiv:2510.17496*, 2025.

[16] Xu Cao. Enhancing video vision language model with hippocampal sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 33682–33692, 2026.

[17] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M Rehg, et al. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21819–21830. IEEE, 2024.

[18] Xu Cao, Yifan Shen, Bolin Lai, Wenqian Ye, Yunsheng Ma, Joerg Heintz, Jintai Chen, Meihuan Huang, Jianguo Cao, Aidong Zhang, and James Matthew Rehg. What is the visual cognition gap between humans and multimodal LLMs? In *Second Conference on Language Modeling*, 2025.

[19] Xu Cao, Pranav Virupaksha, Wenqi Jia, Bolin Lai, Fiona Ryan, Sangmin Lee, and James M Rehg. Socialgesture: Delving into multi-person gesture understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19509–19519. IEEE, 2025.

[20] Xu Cao, Pranav Virupaksha, Sangmin Lee, Bolin Lai, Wenqi Jia, Jintai Chen, and James Rehg. Toward human deictic gesture target estimation. *Advances in Neural Information Processing Systems*, 38:172399–172424, 2025.

[21] Xu Cao, Houze Yang, Vipin Gunda, Zhongyi Zhou, Tianyu Xu, Adarsh Kowdle, Inki Kim, and James M Rehg. Gaze target estimation anywhere with concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 31304–31315, 2026.

[22] Miguel Carvalho and Bruno Martins. Efficient architectures for high resolution vision-language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10520–10530, 2025.

[23] Raymond B Cattell. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1):1–22, 1963.

[24] Liwei Che, Yihao Quan, Sen Fang, Hongyi Wang, Ranjay Krishna, Ruixiang Tang, and Vladimir Pavlovic. Why MLLMs struggle to count: Overcoming individuation and aggregation bottlenecks with ConvStack. *arXiv preprint arXiv:2609.37096*, 2026.

[25] 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 lite vision-language models. *arXiv preprint arXiv:2402.11684*, 2024.

[26] Liang Chen, Weichu Xie, Yiyan Liang, Hongfeng He, Hans Zhao, Zhibo Yang, Zhiqi Huang, Haoning Wu, Haoyu Lu, Y. Charles, Yiping Bao, et al. BabyVision: Visual reasoning beyond language. *arXiv preprint arXiv:2601.06521*, 2026.

[27] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024.

[28] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

[29] Yew Ken Chia, Vernon Toh, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16259–16273, 2024.

[30] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

[31] Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. ARC-AGI-2: A new challenge for frontier AI reasoning systems. *arXiv preprint arXiv:2505.11831*, 2025.

[32] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[33] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36: 49250–49267, 2023.

[34] Sen Fang, Chen Chen, Lei Wang, Ce Zheng, Chunyu Sui, and Yapeng Tian. Signllm: Sign language production large language models. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 6681–6693. IEEE, 2025.

[35] Sen Fang, Yalin Feng, Chunyu Sui, Hongbin Zhong, Yanxin Zhang, Hongwei Yi, Hezhen Hu, and Dimitris N Metaxas. SignX: Continuous sign recognition in compact pose-rich latent space. *arXiv preprint arXiv:2504.16315*, 2025.

[36] Sen Fang, Chunyu Sui, Yanghao Zhou, Xuedong Zhang, Hongbin Zhong, Yapeng Tian, and Chen Chen. Signdiff: Diffusion model for american sign language production. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–11. IEEE, 2025.

[37] Sen Fang, Hongbin Zhong, Yalin Feng, Yanxin Zhang, and Dimitris N Metaxas. Streamflow: Theory, algorithm, and implementation for high-efficiency rectified flow generation. *arXiv preprint arXiv:2511.22009*, 2025.

[38] Sen Fang, Yalin Feng, Yanxin Zhang, and Dimitris N Metaxas. RAC: Rectified flow auto coder. *arXiv preprint arXiv:2603.05925*, 2026.

[39] Sen Fang, Yalin Feng, Hongbin Zhong, Yanxin Zhang, and Dimitris N Metaxas. Stable signer: Hierarchical sign language generative model. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14472–14483, 2026.

[40] Sen Fang, Hongbin Zhong, Yanxin Zhang, and Dimitris N Metaxas. SignVerse-2M: A two-million-clip pose-native universe of 55+ sign languages. *arXiv preprint arXiv:2605.01720*, 2026.

[41] François Fleuret, Ting Li, Charles Dubout, Emma K Wampler, Steven Yantis, and Donald Geman. Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, 108(43):17621–17625, 2011.

[42] 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*, pages 148–166. Springer, 2024.

[43] Giorgio Ganis and Haline E Schendan. Visual imagery. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):239–252, 2011.

[44] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[45] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7 (2):155–170, 1983.

[46] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.

[47] Qing Guo, Prashan Wanigasekara, Jian Zheng, Jacob Zhiyuan Fang, Xinwei Deng, and Chenyang Tao. How do large multimodal models really fare in classical vision few-shot challenges? A deep dive. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

[48] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023.

[49] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can MLLMs reason in multimodality? EMMA: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025.

[50] Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. Stratified rule-aware network for abstract visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1567–1574, 2021.

[51] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024.

[52] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023.

[53] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019.

[54] Susanne M Jaeggi, Barbara Studer-Luethi, Martin Buschkuehl, Yi-Fen Su, John Jonides, and Walter J Perrig. The relationship between n-back performance and matrix reasoning—implications for training and transfer. *Intelligence*, 38(6):625–635, 2010.

[55] Wenqi Jia, Bolin Lai, Xu Cao, Miao Liu, Danfei Xu, and James M Rehg. Learning predictive visuomotor coordination. In *Findings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3609–3619, 2026.

[56] Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, and Anima Anandkumar. Bongard-HOI: Benchmarking few-shot visual reasoning for human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19056–19065, 2022.

[57] Alan S Kaufman, Susan Engi Raiford, and Diane L Coalson. *Intelligent testing with the WISC-V*. John Wiley & Sons, 2015.

[58] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pages 235–251. Springer, 2016.

[59] Salem Lahlou, Tristan Deleu, Pablo Lemos, Dinghuai Zhang, Alexandra Volokhova, Alex Hernández-Garcıa, Léna Néhale Ezzine, Yoshua Bengio, and Nikolay Malkin. A theory of continuous generative flow networks. In *International Conference on Machine Learning*, pages 18269–18300. PMLR, 2023.

[60] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

[61] Sangmin Lee, Minzhi Li, Bolin Lai, Wenqi Jia, Fiona Ryan, Xu Cao, Ozgur Kara, Bikram Boote, Weiyan Shi, Diyi Yang, et al. Towards social AI: A survey on understanding social interactions. *arXiv preprint arXiv:2409.15316*, 2024.

[62] Solim LeGris, Wai Keen Vong, Brenden M Lake, and Todd M Gureckis. H-ARC: a robust estimate of human performance on the abstraction and reasoning corpus benchmark. *arXiv preprint arXiv:2409.01374*, 2024.

[63] Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025.

[64] Bingxuan Li, Yiwei Wang, Tao Meng, Kai-Wei Chang, and Nanyun Peng. Control large language models via divide and conquer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15240–15256, 2024.

[65] Bingxuan Li, Yiwei Wang, Jiuxiang Gu, Kai-Wei Chang, and Nanyun Peng. METAL: A multi-agent framework for chart generation with test-time scaling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30054–30069, 2025.

[66] Bingxuan Li, Yiming Cui, Yicheng He, Yiwei Wang, Shu Zhang, Longyin Wen, and Yulei Niu. Echofoley: Event-centric hierarchical control for video grounded creative sound generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27229–27238, 2026.

[67] Bingxuan Li, Simo Du, and Yue Guo. Joint optimization of reasoning and dual-memory for self-learning diagnostic agent. *arXiv preprint arXiv:2604.07269*, 2026.

[68] Bingxuan Li, Yining Hong, Cheng Qian, Hyeonjeong Ha, Jiateng Liu, Zhenhailong Wang, Yue Guo, Yunzhu Li, and Heng Ji. Trimming the long-tail of visual world modeling evaluation. *arXiv preprint arXiv:2606.24256*, 2026.

[69] Bingxuan Li, Jeonghwan Kim, Cheng Qian, Xiusi Chen, Eitan Anzenberg, Niran Kundapur, and Heng Ji. Pearl: Self-evolving assistant for time management with reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33522–33539, 2026.

[70] Bingxuan Li, Pengyi Shi, and Amy R Ward. Enhancing predictive model learning via domain-knowledge augmented latent feature mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38853–38861, 2026.

[71] Bingxuan Li, Siqi Song, Yizhuo Wu, Jiarui Yao, Tong Zhang, and Huan Zhang. MotorMind: Scaffolding general vision language models for zero-shot robot manipulation. *arXiv preprint arXiv:2609.38078*, 2026.

[72] Bingxuan Li, Yiwen Song, Xueqing Wu, Yanzhou Pan, Yang Li, Kuang Su, Jingyun Liu, Sebastian Ko, Huan Zhang, Tong Zhang, et al. VIS-Ground: Video interactive storytelling with contextual grounding. *arXiv preprint arXiv:2610.09326*, 2026.

[73] Bingxuan Li, Rui Yang, Cheng Qian, Jiateng Liu, Jeonghwan Kim, Zhenhailong Wang, Manling Li, Tong Zhang, and Heng Ji. Long-horizon embodied decision-making via multimodal memory compression. *arXiv preprint arXiv:2608.01456*, 2026.

[74] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025.

[75] Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025.

[76] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023.

[77] Kelvin Li, Chuyi Shang, Leonid Karlinsky, Rogerio Feris, Trevor Darrell, and Roei Herzig. Latent implicit visual reasoning. *arXiv preprint arXiv:2512.21218*, 2025.

[78] Kevin Y Li, Sachin Goyal, Joao D Semedo, and J Zico Kolter. Inference optimal VLMs need fewer visual tokens and more parameters. *arXiv preprint arXiv:2411.03312*, 2024.

[79] Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. *arXiv preprint arXiv:2506.04633*, 2025.

[80] Xinzhuo Li, Adheesh Juvekar, Jiaxun Zhang, Xingyou Liu, Muntasir Wahed, Kiet A Nguyen, Yifan Shen, Tianjiao Yu, and Ismini Lourentzou. Counterfactual segmentation reasoning: Diagnosing and mitigating pixel-grounding hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7450–7460, 2026.

[81] Zirui Li, Xuefeng Bai, Kehai Chen, Yizhi Li, Jian Yang, Chenghua Lin, and Min Zhang. Dynamics within latent chain-of-thought: An empirical study of causal structure. *arXiv preprint arXiv:2602.08783*, 2026.

[82] Jia Liang and Liangming Pan. Do Latent-CoT models think step-by-step? A mechanistic study on sequential reasoning tasks. *arXiv preprint arXiv:2602.00449*, 2026.

[83] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916, 2023.

[84] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[85] 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*, pages 216–233. Springer, 2024.

[86] Yuanzhe Liu, Jingyuan Zhu, Yuchen Mo, Gen Li, Xu Cao, Jin Jin, Yifan Shen, Zhengyuan Li, Tianjiao Yu, Wenzhen Yuan, Fangqiang Ding, and Ismini Lourentzou. Palm: Progress-aware policy learning via affordance reasoning for long-horizon robotic manipulation. *arXiv preprint arXiv:2601.07060*, 2026.

[87] Yutong Liu, Nan Huang, Xu Cao, and James M Rehg. Thinking in pictures: A systematic benchmark for reasoning-driven image generation. *arXiv preprint arXiv:2609.02864*, 2026.

[88] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[89] 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 *The Twelfth International Conference on Learning Representations*, 2024.

[90] Kanika Madan, Jarrid Rector-Brooks, Maksym Korablyov, Emmanuel Bengio, Moksh Jain, Andrei Cristian Nica, Tom Bosc, Yoshua Bengio, and Nikolay Malkin. Learning gflownets from partial episodes for improved convergence and stability. In *International Conference on Machine Learning*, pages 23467–23483. PMLR, 2023.

[91] Mikołaj Małkiński and Jacek Mańdziuk. One self-configurable model to solve many abstract visual reasoning problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14297–14305, 2024.

[92] Mikołaj Małkiński and Jacek Mańdziuk. A-I-RAVEN and I-RAVEN-Mesh: Two new benchmarks for abstract visual reasoning. *arXiv preprint arXiv:2406.11061*, 2024.

[93] Mikołaj Małkiński and Jacek Mańdziuk. Deep learning methods for abstract visual reasoning: A survey on Raven's Progressive Matrices. *ACM Computing Surveys*, 57(7):1–36, 2025.

[94] Mikołaj Małkiński, Szymon Pawlonka, and Jacek Mańdziuk. Reasoning limitations of multimodal large language models. A case study of Bongard problems. *arXiv preprint arXiv:2411.01173*, 2024.

[95] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, 2022.

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

[97] Meta. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. Online, 2025. URL `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`.

[98] MistralAI. Introducing mistral 3. Online, 2025. URL `https://mistral.ai/news/mistral-3`.

[99] Kiet A Nguyen, Adheesh Juvekar, Tianjiao Yu, Muntasir Wahed, and Ismini Lourentzou. Calico: Part-focused semantic co-segmentation with large vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4550–4561. IEEE, 2025.

[100] Weili Nie, Zhiding Yu, Lei Mao, Ankit B. Patel, Yuke Zhu, and Animashree Anandkumar. Bongard-LOGO: A new benchmark for human-level concept learning and reasoning. In *Advances in Neural Information Processing Systems*, volume 33, pages 16468–16480, 2020.

[101] OpenAI. Introducing GPT-5.2. Online, 2025. URL `https://openai.com/index/introducing-gpt-5-2/`.

[102] OpenAI. Introducing OpenAI o3 and o4-mini. Online, 2025. URL `https://openai.com/index/introducing-o3-and-o4-mini/`.

[103] Szymon Pawlonka, Mikołaj Małkiński, and Jacek Mańdziuk. Bongard-RWR+: Real-world representations of fine-grained concepts in Bongard problems. *arXiv preprint arXiv:2508.12026*, 2025.

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

[105] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmLR, 2021.

[106] John Raven and Jean Raven. Raven progressive matrices. In *Handbook of nonverbal assessment*, pages 223–237. Springer, 2003.

[107] John Raven, John C. Raven, and John H. Court. *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Oxford Psychologists Press, Oxford, England, 1998.

[108] John C. Raven. Mental tests used in genetic studies: The performance of related individuals on tests mainly educative and mainly reproductive. Master's thesis, University of London, 1936.

[109] Eleanor H Rosch. Natural categories. *Cognitive Psychology*, 4(3):328–350, 1973.

[110] Timothy A Salthouse. Influence of working memory on adult age differences in matrix reasoning. *British Journal of Psychology*, 84(2):171–199, 1993.

[111] Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7(1):96–106, 2025.

[112] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

[113] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in VLMs. *Advances in Neural Information Processing Systems*, 38:17929–17960, 2025.

[114] Yifan Shen, Chuanmiao Dong, Zhuoqing Zhong, Pei Tian, Tianjiao Yu, Jiateng Liu, Bowen Fang, Xinzhuo Li, Yuanzhe Liu, Zhengyuan Li, et al. Position: Multimodal LLMs should learn from children. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 5619–5630, 2026.

[115] Yifan Shen, Jiateng Liu, Xinzhuo Li, Yuanzhe Liu, Bingxuan Li, Houze Yang, Wenqi Jia, Yijiang Li, Tianjiao Yu, James Matthew Rehg, Xu Cao, and Ismini Lourentzou. EgoForge: Goal-directed egocentric world simulator. *arXiv preprint arXiv:2603.20169*, 2026.

[116] Yifan Shen, Pei Tian, Xinzhuo Li, Bowen Fang, Shujun Xia, Bingxuan Li, Ana Jojic, Wenming Ye, Xu Cao, James Matthew Rehg, and Ismini Lourentzou. CogniRoute: Learning to route social evidence in omni-modal models. *arXiv preprint arXiv:2606.20970*, 2026.

[117] Yifan Shen, Jian Xu, Boyi Li, Yuner Zhang, Tianjiao Yu, Bingxuan Li, Houze Yang, Rushi Wang, and Xu Cao. ChronoVision: Temporal reasoning via latent state reconstruction. *arXiv preprint arXiv:2608.05631*, 2026.

[118] Yifan Shen, Jiawen Zhang, Jian Xu, Junho Kim, Ismini Lourentzou, Xu Cao, and Meihuan Huang. Evaluating cognitive age alignment in interactive AI agents. *arXiv preprint arXiv:2605.17894*, 2026.

[119] Ying Shen, Jerry Xiong, Tianjiao Yu, and Ismini Lourentzou. Phantom: Physics-infused video generation via joint modeling of visual and latent physical dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[120] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[121] Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025.

[122] Guohao Sun, Hang Hua, Jian Wang, Jiebo Luo, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Latent chain-of-thought for visual reasoning. *arXiv preprint arXiv:2510.23925*, 2025.

[123] Onkar Susladkar, Tushar Prakash, Gayatri Deshmukh, Kiet A Nguyen, Jiaxun Zhang, Adheesh Juvekar, Tianshu Bao, Lin Chai, Sparsh Mittal, Inderjit S Dhillon, and Ismini Lourentzou. Best of both worlds: Multimodal reasoning and generation via unified discrete flow matching. In *International Conference on Machine Learning*, 2026.

[124] Afrina Tabassum, Bin Guo, Xiyao Ma, Hoda Eldardiry, and Ismini Lourentzou. Mmplanner: Zero-shot multimodal procedural planning with chain-of-thought object state reasoning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025.

[125] Younes Adam Tabi, Maria Raquel Maio, Bahaaeddin Attaallah, Shannon Dickson, Daniel Drew, Mohamad Imran Idris, Annika Kienast, Verena Klar, Lisa Nobis, Olivia Plant, et al. Vividness of visual imagery questionnaire scores and their relationship to visual short-term memory performance. *Cortex*, 146:186–199, 2022.

[126] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini Robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[127] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[128] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9568–9578, 2024.

[129] Shimon Ullman. Visual routines. In *Readings in computer vision*, pages 298–328. Elsevier, 1987.

[130] Muntasir Wahed, Kiet A Nguyen, Adheesh Sunil Juvekar, Xinzhuo Li, Xiaona Zhou, Vedant Shah, Tianjiao Yu, Pinar Yanardag, and Ismini Lourentzou. Prima: Multi-image vision-language models for reasoning segmentation. *arXiv preprint arXiv:2412.15209*, 2024.

[131] Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. *arXiv preprint arXiv:2410.09575*, 2024.

[132] Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. MindCube: Spatial mental modeling from limited views. In *The Fourteenth International Conference on Learning Representations*, 2026.

[133] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.

[134] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[135] Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. Site: towards spatial intelligence thorough evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9058–9069, 2025.

[136] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[137] David Wechsler. *Wechsler intelligence scale for children.* Psychological corporation, 1949.

[138] David Wechsler. *WAIS-IV Administration and Scoring Manual*. PsychCorp, 2008.

[139] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[140] Max Wertheimer. Untersuchungen zur Lehre von der Gestalt. II. *Psychologische Forschung*, 4 (1):301–350, 1923.

[141] Rujie Wu, Xiaojian Ma, Zhenliang Zhang, Wei Wang, Qing Li, Song-Chun Zhu, and Yizhou Wang. Bongard-OpenWorld: Few-shot reasoning for free-form visual concepts in the real world. In *International Conference on Learning Representations*, 2024.

[142] xAI. Grok-1.5 vision preview, 2024. URL `https://x.ai/blog/grok-1.5v`.

[143] xAI. Grok 4 fast model card. System card (technical report), September 2025. URL `https://data.x.ai/2025-09-19-grok-4-fast-model-card.pdf`.

[144] Siying Xie, Daniel Kaiser, and Radoslaw M Cichy. Visual imagery and perception share neural representations in the alpha frequency band. *Current Biology*, 30(13):2621–2627, 2020.

[145] Jingyi Xu, Tushar Vaidya, Yufei Wu, Saket Chandra, Zhangsheng Lai, and Kai Fong Ernest Chong. Abstract visual reasoning: An algebraic approach for solving Raven's Progressive Matrices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6715–6724, 2023.

[146] Yifan Xu, Yixuan Li, Xinzhuo Li, Yixin Gu, Yifan Shen, Lijun Yu, and Haohan Wang. STEVE: Stabilizing textual gradient-based prompt optimization via error-driven refinement and regularized verification. *arXiv preprint arXiv:2609.23716*, 2026.

[147] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.

[148] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L Brown II, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-S: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*, 2026.

[149] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025.

[150] Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025.

[151] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[152] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

[153] Eunice Yiu, Maan Qraitem, Anisa Noor Majhi, Charlie Wong, Yutong Bai, Shiry Ginosar, Alison Gopnik, and Kate Saenko. KiVA: Kid-inspired visual analogies for testing large multimodal models. In *International Conference on Learning Representations*, 2025.

[154] Tianjiao Yu, Xinzhuo Li, Yifan Shen, Yuanzhe Liu, and Ismini Lourentzou. CoRe3D: Collaborative reasoning as a foundation for 3D intelligence. *arXiv preprint arXiv:2512.12768*, 2025.

[155] Tianjiao Yu, Xinzhuo Li, Yifan Shen, Ying Shen, Kiet A Nguyen, Adheesh Sunil Juvekar, and Ismini Lourentzou. SILSA: Sliding-window slice latents for topology-preserving high-resolution 3D generation. *arXiv preprint arXiv:2610.02201*, 2026.

[156] Tianjiao Yu, Xinzhuo Li, Yifan Shen, Onkar Susladkar, Yuanzhe Liu, Xiaona Zhou, and Ismini Lourentzou. ELSA3D: Elastic semantic anchoring for unified 3D understanding and generation. *arXiv preprint arXiv:2607.06565*, 2026.

[157] Tianjiao Yu, Xinzhuo Li, Muntasir Wahed, Jerry Xiong, Yifan Shen, Ying Shen, and Ismini Lourentzou. Dreampartgen: Semantically grounded part-level 3d generation via collaborative latent denoising. In *European Conference on Computer Vision*, pages 94–112. Springer, 2026.

[158] 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*, pages 9556–9567, 2024.

[159] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, 2025.

[160] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

[161] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019.

[162] Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A benchmark for compositional visual reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 29776–29788, 2022.

[163] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. RAVEN: A dataset for relational and analogical visual rEasoNing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5317–5327, 2019.

[164] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. MM-LLMs: Recent advances in multimodal large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12401–12430, 2024.

[165] Haichao Zhang and Yun Fu. Vqtoken: Neural discrete token representation learning for extreme token reduction in video large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[166] Haichao Zhang, Can Qin, Yu Yin, and Yun Fu. Camouflaged image synthesis is all you need to boost camouflaged detection. *arXiv preprint arXiv:2308.06701*, 2023.

[167] Haichao Zhang, Yi Xu, Hongsheng Lu, Takayuki Shimizu, and Yun Fu. Layout sequence prediction from noisy mobile modality. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3965–3974, 2023.

[168] Haichao Zhang, Yi Xu, Hongsheng Lu, Takayuki Shimizu, and Yun Fu. Oostraj: Out-of-sight trajectory prediction with vision-positioning denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14802–14811, 2024.

[169] Haichao Zhang, Wenhao Chai, Shwai He, Ang Li, and Yun Fu. DIVE-Bench: Codec-style high-FPS video understanding. In *European Conference on Computer Vision*, pages 418–435. Springer, 2026.

[170] Haichao Zhang, Mingfei Chen, Shwai He, Zhengtong Xu, Yifan Shen, Yiyang Huang, Jianglin Lu, Yijiang Li, Yu She, and Yun Fu. A survey of physical ai: A history from chatgpt to world models and embodied agents. *Preprints, June*, 2026.

[171] Haichao Zhang, Yijiang Li, Shwai He, Tushar Nagarajan, Mingfei Chen, Jianglin Lu, Ang Li, and Yun Fu. Thinkjepa: Empowering latent world models with large vision-language reasoning model. *arXiv preprint arXiv:2603.22281*, 2026.

[172] Haichao Zhang, Yao Lu, Lichen Wang, Yunzhe Li, Daiwei Chen, Yunpeng Xu, and Yun Fu. LinkedOut: Linking world knowledge representation out of video LLM for next-generation video recommendation. In *Findings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7111–7121, 2026.

[173] Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, et al. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in MLLMs. *arXiv preprint arXiv:2510.24514*, 2025.

[174] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024.

[175] Yizhe Zhang, He Bai, Ruixiang Zhang, Jiatao Gu, Shuangfei Zhai, Josh Susskind, and Navdeep Jaitly. How far are we from intelligent visual deductive reasoning? *arXiv preprint arXiv:2403.04732*, 2024.

[176] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[177] Xiaona Zhou, Muntasir Wahed, Tianjiao Yu, Constantin Brif, and Ismini Lourentzou. Tiny but trusted: Efficient vision-language reasoning for time-series anomaly detection. *arXiv preprint arXiv:2605.30344*, 2026.

[178] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[179] Rui-Jie Zhu, Tianhao Peng, Tianhao Cheng, Xingwei Qu, Jinfa Huang, Dawei Zhu, Hao Wang, Kaiwen Xue, Xuanliang Zhang, Yong Shan, et al. A survey on latent reasoning. *arXiv preprint arXiv:2507.06203*, 2025.

Appendix

A Related Works

Bongard Problems. The Bongard Problems (BPs) were introduced by M. M. Bongard [12, 13] as a novel challenge to machine vision. A typical BP includes two sets, positive and negative, with each side consisting of six images that share a common pattern. It requires a system to induce the logical rule that distinguishes a set of positive examples from negative ones [12, 13]. This task evaluates core properties of human cognition: identifying the underlying rule that differentiates the two sides and articulating it in natural language [94]. To modernize this benchmark for deep learning, the Bongard-LOGO dataset was introduced, transforming the concept learning challenge into a few-shot binary classification problem [100]. Recent works have expanded the field of BPs to include real-world images, such as Bongard-HOI [56], Bongard-OpenWorld [141], Bongard-RWR [94], and Bongard-RWR+ [103]. These works expand the range of presented objects, attributes, and relationships, illustrating human-object interactions and incorporating free-form concepts in the real world, thus increasing the diversity of featured scenes.

Matrix Reasoning. Matrix reasoning tasks play a critical role in assessing human intelligence, particularly in relation to visual cognition and working memory [110, 54, 41]. These tasks are widely used in psychology through Raven's Progressive Matrices (RPMs) [108, 107, 106], the Wechsler Intelligence Scale for Children (WISC) [137, 57], and the Wechsler Adult Intelligence Scale (WAIS-IV) [138] to evaluate human fluid intelligence and abstract visual reasoning. Early works, such as PGM [7] and RAVEN [163], were proposed to test whether neural networks can learn abstract reasoning and to lift machine intelligence by associating vision with structural, relational, and analogical reasoning in a hierarchical representation. Subsequent works further demonstrated that deep learning models can be trained to solve simple matrix reasoning tasks [93, 145, 91]. Several datasets and benchmarks have also been introduced to enlarge this field, such as I-RAVEN [50], RAVEN-FAIR [10], and CVR [162]. However, these studies tend to overlook the ability of humans to solve such puzzles in a zero-shot manner, without the need for explicit training on extensive datasets. Therefore, some useful zero-shot visual reasoning inference datasets such as RAVEN-IQ [52] and Visual Reasoning Benchmark [175] have been proposed to deal with the issue. The latest MaRs-VQA [18] addresses the limitations of previous works, namely the lack of rigorous human experiments as a reference and the reliance on relatively small datasets without psychometric validation.

Abstraction and Reasoning Corpus. Abstraction and Reasoning Corpus for Artificial General Intelligence (ARC-AGI) is a novel general AI benchmark designed to measure a human-like form of general fluid intelligence and enable fair general intelligence comparisons between AI systems and humans [30]. Its measurement standard emphasizes skill-acquisition efficiency and out-of-domain generalization over mere task performance. The tasks require systems to solve grid-based visual program synthesis problems by abstracting the underlying rules from only a minimal set of input-output examples (typically three to five pairs). A robust estimate of human performance [62] on the full ARC public evaluation set places the average human accuracy at 64.2%, far exceeding that of current state-of-the-art AI methods. Its updated version, ARC-AGI-2 [31], is more challenging. Even the most powerful models at the time of its release, such as OpenAI's o3 [102], achieve only negligible performance on the newly released, adversarially selected ARC-AGI-2 challenge. This reinforces that modern frontier reasoning models fundamentally lack the structural inductive biases that are necessary for human-like skill-acquisition and generalization.

VLMs for Vision Reasoning. Visual recognition-related Vision Language Model (VLM) studies have made great progress since the development of CLIP [105]. Prior work has shown that VLMs are capable of addressing vision reasoning tasks [161, 14, 113, 86, 72, 65, 64, 172, 99, 130, 80, 177]. VLMs are the dominant architecture for bridging perception and language, typically integrating a visual encoder (*e.g.*, CLIP) with pretrained LLMs via a cross-modal connector to align visual features with text space [83, 17, 164, 112, 42, 165, 169]. The training pipeline for VLMs often involves pre-training for modality alignment, followed by instruction tuning or Supervised Fine-Tuning (SFT) to enhance general multimodal capabilities [83, 151]. Models such as Qwen-VL [3, 6, 5], LLaVA [83, 84], MiniGPT-4 [178], InstructBLIP [33], and CogVLM [133] highlight the significance of employing high-quality visual instruction tuning data. However, current VLMs still face challenges in adapting to high-resolution [22, 78] and visually complex images in vision reasoning. This is because most current vision reasoning approaches are primarily text-level, with the LLM

Table B1: **CogSense Benchmark Dimensions and Psychometric Constructs.** We define five CogSense dimensions based on established psychometric constructs and cognitive theories, with each dimension comprising multiple cognitive subsets.

Dimension	Psychometric Construct	Source	Subsets
Fluid Intelligence	G_{f}	Theory of Fluid and Crystallized Intelligence [23], Structure-mapping [45]	MaRs-VQA, PGM, RAVEN
Crystallized Intelligence	G_{c}	Theory of Fluid and Crystallized Intelligence [23], Natural Categories [109]	Bongard-RWR+, Bongard-HOI
Visuospatial Cognition	G_{v}	Untersuchungen zur Lehre von der Gestalt [140], Recognition-by-components [11]	Bongard-LOGO
Mental Simulation	$G_{\mathrm{v}}/G_{\mathrm{f}} \cap$ Causal Reasoning	Simulation as an Engine of Physical Scene Understanding [8]	KiVA, STARE, ARC-AGI, ARC-AGI-2
Visual Routines	$G_{\mathrm{v}}/G_{\mathrm{f}} \cap$ Sequential Attention	Visual Routines [129]	CVR

exploring textual tokens while the visual input remains static [63]. This absence of a dynamic visual search mechanism limits the model's ability to acquire fine-grained visual cues selectively. Other limitations stem from few-shot reasoning [47, 70], compositional understanding [160, 24], the constrained visual grounding capabilities inherent in CLIP [128, 166], the brittleness of textual gradient-based optimization [146], and the challenge of dynamically integrating multimodal evidence [116, 73, 67, 69]. Our work addresses these problems, enabling high-performing vision reasoning via step-by-step latent visual imagery prediction.

Latent Visual Reasoning. While standard VLMs rely on static visual encoding, a new paradigm of *Latent Visual Reasoning* has emerged, shifting the focus from passive perception to active, internal simulation. This approach draws inspiration from World Models [179, 122, 171, 170, 68], enabling systems to "think" in a compressed latent space before generating a response. Recent advancements extend this paradigm to predict dynamic changes via temporal latent reconstruction [117, 168, 167] and embed reasoning within goal-directed egocentric simulators [115, 71]. Rather than mapping pixels directly to text, these models perform intermediate reasoning steps, akin to System-2 cognition, by predicting future states or manipulating visual abstractions [179, 122, 55]. For instance, recent frameworks introduce "visual scratchpads" or latent tokens that allow the model to sketch out reasoning traces implicitly [173, 74]. Methods such as Mirage [150] and Latent Sketchpad [173] exemplify this by empowering models to generate and refine mental imagery during the inference process, effectively bridging the gap between visual grounding and abstract logic. Further advancements include Latent Visual Reasoning (LVR) [63] and implicit reasoning tokens [77], which allow models to perform autoregressive planning in the visual embedding space without explicit supervision. To effectively untangle complex visual environments, approaches like ROSS [131] complement this paradigm by explicitly modeling spatial topological structures within the latent space via a denoising objective to reconstruct latent representations of the question image as vision-centric supervision, augmenting the model's structural comprehension. These methods overcome the limitations of text-centric reasoning by maintaining rich, high-dimensional visual information throughout the decision-making chain [63, 77]. Furthermore, similar principles of latent semantic grounding have also driven advances in continuous 3D spatial modeling [154, 156, 157, 155] and generative modeling [34, 36, 39, 35, 40, 37, 38, 66]. While these prior works demonstrate the potential of utilizing latent spaces for intermediate reasoning, they mainly rely on either unconstrained autoregressive planning or vision-centric supervision derived from the input context. In contrast, our work predicts visual hypotheses conditioned on intermediate reasoning prefixes, and the latent-rationale optimization stage uses their consistency to shape the reasoning policy.

B Theoretical Grounding of the Five Cognitive Dimensions

As shown in Table B1, the first three dimensions represent the traditional psychometric approach, while the last two dimensions belong to the process-level tradition of visual cognition. Together, these dimensions cover complementary operations required for high-level visual cognition: inducing rules, retrieving concepts, organizing spatial structure, simulating changes, and executing attention-based visual operations. We exclude other traditional psychometric dimensions such as Working Memory (G_{wm}) and Processing Speed (G_{s}). These dimensions measure cognitive capacity and execution speed rather than reasoning operations. In the context of MLLMs, they are largely confounded with

Table C1: **CogSense-Bench descriptions and statistics.** CogSense-Bench comprises multiple datasets grouped into five cognitive categories, covering diverse aspects of visual reasoning, including fluid intelligence, crystallized intelligence, visuospatial cognition, mental simulation, and visual routines. We report the number of training and test VQA pairs. Dataset licenses are provided when available.

Category	Dataset	# Train	# Test	License
Fluid Intelligence	MaRs-VQA	1.4K	13	CC-BY-NC-3.0
	PGM	10K	94	Unspecified
	RAVEN	18K	169	GPL-3.0 License
Crystallized Intelligence	Bongard-RWR+	16K	152	CC-BY-4.0
	Bongard-HOI	23K	216	NVIDIA Source Code License
Visuospatial Cognition	Bongard-LOGO	12K	113	MIT license
Mental Simulation	KiVA	1.4K	13	Apache-2.0 License
	STARE	4K	26	Unspecified
	ARC-AGI	1.6K	31	Apache-2.0 License
	ARC-AGI-2	8K	80	Apache-2.0 License
Visual Routines	CVR	10K	93	Apache-2.0 License
Sum		105.4K	1000	

architectural hyperparameters (*e.g.*, context window length and KV-cache) rather than inner visual cognitive capabilities.

C Statistics of CogSense-Bench

We summarize the statistics of CogSense-Bench in Table C1. Throughout the paper, Avg. denotes the overall accuracy over all 1,000 test items, i.e., the category accuracies weighted by the number of test items in each category, rather than the unweighted mean of the five categories. We ensured zero data leakage by randomly sampling instances from each category at the same proportion as in the train split and subsequently removing the sampled data from it.

D Implementation Details for Data Pipeline

D.1 Data Extraction

We curate visual cognition data from a broad set of datasets spanning Bongard-style reasoning, abstract visual reasoning, spatial cognition, and ARC-style program induction, including Bongard-OpenWorld [141], Bongard-HOI [56], Bongard-LOGO [100], Bongard-RWR+ [103], CVR [162], KiVA [153], MaRs-VQA [18], PGM [7], RAVEN [163], I-RAVEN [50], RAVEN-FAIR [10], A-I-RAVEN [92], I-RAVEN-Mesh [92], I-RAVEN-X [15], STARE [79], ARC-AGI [30], and ARC-AGI-2 [31]. From these sources, we manually select examples that match our five predefined cognitive categories, ensuring that each retained instance is relevant to the target reasoning skills evaluated in our benchmark. Building on these seed sources, we develop automated extraction and cleaning scripts to expand the collection at scale and produce the raw dataset. To standardize evaluation across heterogeneous task formats, we further convert tasks that were not originally multiple-choice into a unified multiple-choice format. For Bongard-style problems, we randomly select one image from the positive set as the correct answer, combine it with negative samples, and shuffle the resulting candidates as answer options; the remaining positive samples are used to define the query context. For ARC-AGI-style problems, we perform data augmentation on the ground-truth images and present both the original ground-truth image and augmented variants as candidate options.

D.2 Reasoning Chain Generation

We transform traditional short QA pairs into high-quality reasoning data containing an explicit reasoning chain. We customize different prompts for different types of original questions and leverage LLMs to generate corresponding reasoning chains based on these different prompts. The teacher model $\mathcal{M}_T$ is required not only to output the final answer but also to elaborate on its reasoning process. Specifically, it is guided to analyze visual cues within the images, infer underlying rules, and explain causal relationships, finally generating detailed reasoning process text. To ensure quality,

Reasoning Chain Generation Prompts for CogSense-Bench

You are a careful visual reasoning assistant. Your task is to generate CoT and the final answer for the provided questions.
Your response must consist of **ONLY** the following two parts in JSON format:

1. 'reasoning': reasoning steps (3-12 concise steps from high-level to details).
2. 'final_answer': the final answer of the question. Keep reasoning concise (3-12 steps). **Do NOT** include internal deliberation beyond the steps.

Output must be in JSON format and **ONLY** with the reasoning steps and final answer with nothing else.

For Crystallized Intelligence and Visuospatial Cognition Questions:
You will be given a set of images that share a common pattern. Analyze these examples to understand the pattern. Then, select the image that belongs to the same category as the images in the question set.
Question images (these images share a common pattern): [question_images]
Option images: [option_images]

For Fluid Intelligence Questions:
You will be given an image that contains eight subimages and a blank space. Then, given some option images, predict which one is the right option, following the same pattern to complete the sequence based on the pattern shown in the question image.
Question image: <question_image>
Option images: [option_images]

For ARC-AGI in Mental Simulation:
You will be given some pairs of input-output images as examples. The output is obtained by coloring the grid according to a specific pattern based on the input. Analyze these examples to understand the underlying grid color pattern. Then, given a new input image and option images, predict which image is generated by applying the same pattern to the input.
Examples: [<input_image, output_image>]
Question image: <new_input_image>
Option images: [option_images]

For KiVA in Mental Simulation:
You will be given some pairs of input-output images or transform patterns as examples. Analyze the pattern and select the correct answer from the given options. Then, given a new input image and some options that indicate possible outputs, predict the right option following the same pattern.
Examples: [pattern_image]
Question image: <new_input_image>
Option images: [option_images]

For STARE in Mental Simulation:
You will be given some images and a blank space. Then, given some option images, predict which one is the right option, following the same pattern to complete the sequence based on the pattern shown in the question image.
Question image: <question_image>
Option images: [option_images]

For Visual Routines Questions:
You will be given some images, in which one of them differs from the rest of the images. Find the odd one out.
Question images: [images]
Option images: [images]

Input: [images] and <prompt>
Output: <reasoning_chain> and <answer>

Figure D1: **Reasoning Chain Generation Prompts.** We customize different prompts for different types of original questions and use LLMs to generate corresponding reasoning chains based on these different prompts.

we employ an automated verification pipeline: (1) The teacher model generates the rationale and a final prediction based on task-specific prompt protocols (Figure D1); (2) We automatically parse the final predicted answer and match it strictly against the ground truth; (3) If the prediction is incorrect,

Table F1: **Performance Comparison under Cambrian.** All models are based on the Qwen3-VL-8B backbone and trained exclusively on the Cambrian-737K dataset.

Model	HallusionBench	MMBench-EN	MMBench-CN	MMMU	AI2D
ROSS (Qwen3-VL-8B, Cambrian)	59.6	80.6	81.2	64.3	79.9
CogSense (Qwen3-VL-8B, Cambrian) *w/o* LVIP	60.7	81.5	81.7	69.6	82.7
CogSense (Qwen3-VL-8B, Cambrian)	**62.5**	**83.7**	**84.1**	**71.1**	**84.2**

the reasoning trace is immediately discarded, and we prompt the teacher model to retry (up to $K = 3$ times); (4) For the subset of highly complex questions where the teacher model consistently failed even after 3 retries, we manually solved them and explicitly authored the step-by-step reasoning traces; (5) For the teacher-generated subset that successfully passed the automated answer-consistency filter, we conducted a human evaluation to ensure the generated rationales do not contain hallucinations. We randomly sampled a subset of 1,000 verified reasoning traces for a blind review by human annotators. The evaluation showed an overall acceptance rate of **95.4%**, demonstrating that our automated answer-consistency filtering highly correlates with rationale quality.

E Human Study Design and Setup

Participant Recruitment. To establish a human performance baseline, we recruited a total of 20 participants through online platforms using a random sampling strategy, with ages ranging from 18 to 34 (average 24.5). They participated purely out of interest in AI cognitive evaluation and to test their own performance against the benchmark. Participation in the study was entirely voluntary, and no financial compensation or monetary incentives were provided to the subjects.

Questionnaire Design. The questionnaire design involved a rigorous selection process to ensure representativeness. We constructed a test set comprising 100 multiple-choice questions derived from the CogSense-Bench. These questions were manually selected by stratified sampling to strictly maintain the category proportions of the original benchmark. The study was implemented and distributed digitally via the Google Forms platform, allowing for remote administration.

Time Spent and Compensation. The survey took participants an average of 45 minutes to complete. Participants proceeded entirely voluntarily and were free to drop out at any point without consequence. To ensure data quality, we monitored response times and confirmed that none of the participants rushed through the survey. The high consistency in their answers also indicates rigorous engagement with the task. Since uncompensated volunteer recruitment limits the representativeness of the human baseline, we acknowledge this limitation and do not present the 20-participant result as a population-level estimate of human cognitive performance.

Consent Procedure. Before evaluation, all participants were presented with a digital informed consent form that outlined the purpose of the study, the estimated time commitment, the voluntary and uncompensated nature of the task, and the data anonymity protocols. Participants were required to explicitly agree to these terms before proceeding. The study protocol, including the consent process and volunteer framework, was reviewed and approved by our Institutional Review Board (IRB) under an existing IRB approval for visual cognition research at our institution.

Compliance. Regarding experimental standards and ethical compliance, all protocols were reviewed and approved by the IRB prior to the study.

F Comparison between ROSS and CogSense

We compare CogSense against ROSS [131] by conducting two more experiments, where we train the same base model Qwen3-VL-8B [5] with both frameworks. For ROSS, we use the same denoiser as [131], freeze Qwen3-VL-8B, and pre-train the denoiser with the same teacher tokenizer under a mixture of 2M caption data (1246K from ShareGPT4V [27] and 707K from ALLaVA [25]). For CogSense, we do not perform any pre-training.

First, we train both models with supervised fine-tuning on Cambrian-737K [127]. For CogSense, we further train the model with the proposed RL method on the same dataset. We evaluate both models on the listed benchmarks. We further ablate our method with and without LVIP under this setting. We

Table F2: **Performance Comparison under CogSense-Bench.** All models are based on the Qwen3-VL-8B backbone and trained exclusively on the train split of CogSense-Bench.

Model	Fluid Intel.	Crystallized Intel.	Visuospatial Cog.	Mental Simu.	Visual Rout.	Avg.
ROSS (Qwen3-VL-8B)	54.0	84.8	63.7	60.7	43.0	66.4
CogSense-8B (Qwen3-VL-8B)	**63.8**	**91.0**	**69.0**	**68.0**	**50.5**	**73.8**

then evaluate this newly trained model across a range of established general-domain and fine-grained multimodal benchmarks like HallusionBench [46], MMBench [85], MMMU [158], and AI2D [58]. In addition, we train Qwen3-VL-8B on ROSS settings with supervised fine-tuning on the train split of CogSense-Bench and then evaluate both models on the test split.

As shown in Table F1 and Table F2, first, under the same Qwen3-VL-8B backbone and Cambrian settings, adding LVIP consistently improves performance across all five general-domain benchmarks. Moreover, under the same setting, Qwen3-VL-8B backbone, and dataset, our framework also outperforms ROSS. These results isolate the effect of LVIP from backbone or data differences, showing that LVIP provides robust gains beyond our cognitive benchmark and transfers to broader multimodal understanding and fine-grained visual reasoning.

G Statistical Significance Analysis

Bootstrap Confidence Intervals. We compute 95% confidence intervals using a stratified paired bootstrap over the benchmark. Each bootstrap sample preserves the category proportions of CogSense-Bench. We use 10,000 bootstrap resamples.

McNemar Test. For pairwise comparison between CogSense-8B and a baseline, we report McNemar's test [96] on paired correctness outcomes over all N benchmark instances. Let b denote the number of instances that CogSense-8B gets correct while the baseline gets wrong, and c denote the number of instances that CogSense-8B gets wrong while the baseline gets correct. The test is conducted on the $n_{\mathrm{d}} = b + c$ discordant pairs. We compute the two-sided exact McNemar p-value using a binomial test under the null hypothesis that both models have equal accuracy, i.e., $b \sim \mathrm{Binomial}(n_{\mathrm{d}}, 0.5)$. The p-value is defined as the probability of observing a value at least as extreme as the observed b under this distribution.

Table G1: **Main Results with 95% CI on CogSense-Bench.** Confidence Intervals are computed by a stratified paired bootstrap over benchmark items.

Model	Accuracy	95% CI
Human	88.4	[85.5, 91.3]
Gemini 2.5 Flash	34.9	[32.0, 37.8]
GPT-o3	32.3	[29.4, 35.2]
GPT-5.2	40.3	[37.4, 43.2]
Claude Sonnet 4	32.6	[29.7, 35.5]
Grok 4 Fast	31.7	[28.8, 34.6]
Llama-4-Scout-17B	31.8	[28.9, 34.7]
Gemma-3-27B	32.7	[29.8, 35.6]
Ministral-3-14B	16.5	[13.6, 19.4]
Intern-S1	17.5	[14.6, 20.4]
Qwen3-VL-30B	37.4	[34.5, 40.3]
InternVL3.5-8B (FT)	63.9	[61.0, 66.8]
Qwen3.5-9B (FT)	69.1	[66.2, 72.0]
Qwen3-VL-8B (Base)	35.5	[32.6, 38.5]
Qwen3-VL-8B SFT *w/o* LVIP	62.3	[59.3, 65.3]
Qwen3-VL-8B SFT *w/* LVIP	68.0	[65.1, 70.8]
Qwen3-VL-8B SFT *w/o* LVIP + GRPO	65.5	[62.6, 68.4]
Qwen3-VL-8B SFT *w/* LVIP + GRPO	70.8	[68.0, 73.6]
Qwen3-VL-8B SFT *w/* LVIP + GRPO + LR	71.5	[68.6, 74.2]
CogSense-8B (Ours)	73.8	[71.0, 76.4]

Our statistical analysis validates the efficacy of our specific training pipeline and architectural choices. As detailed in the paired performance comparisons, CogSense-8B significantly outperforms the strongest same-backbone baseline (Qwen3-VL-8B SFT *w/* LVIP + GRPO + LR, $\Delta = +2.3\%$) under both the stratified paired bootstrap (95% CI: [0.7, 3.9]) and the exact McNemar test ($p = 0.0265$). This supports that the observed improvement is unlikely to be explained by benchmark sampling noise alone, confirming the robustness and effectiveness of our proposed approach.

Table G2: **Paired Performance Comparison Results.** We report bootstrap confidence intervals and McNemar's test results between CogSense-8B and strong baselines.

Comparison (Ours vs. ...)	Δ Acc.	Bootstrap 95% CI	(b, c)	McNemar p
InternVL3.5-8B (FT)	+9.9	[7.4, 12.4]	(145, 46)	$< 10^{-4}$
Qwen3.5-9B (FT)	+4.7	[2.6, 6.8]	(100, 53)	0.0002
Qwen3-VL-8B (Base)	+38.3	[35.2, 41.4]	(420, 37)	$< 10^{-4}$
Qwen3-VL-8B SFT *w/o* LVIP	+11.5	[8.9, 14.1]	(160, 45)	$< 10^{-4}$
Qwen3-VL-8B SFT *w/* LVIP	+5.8	[3.7, 8.1]	(121, 63)	$< 10^{-4}$
Qwen3-VL-8B SFT *w/o* LVIP + GRPO	+8.3	[6.0, 10.6]	(130, 47)	$< 10^{-4}$
Qwen3-VL-8B SFT *w/* LVIP + GRPO	+3.0	[1.2, 4.8]	(78, 48)	0.0095
Qwen3-VL-8B SFT *w/* LVIP + GRPO + LR	+2.3	[0.7, 3.9]	(61, 38)	0.0265

Table I1: **More General Ability Results.** Performance comparison in general ability in vision-language understanding and spatial reasoning benchmarks with CogSense-8B and the base model. CogSense-8B maintains robust performance on these general benchmarks, comparable to the base model.

Model	AI2D	ChartQA	MMStar	MMMU	MMMU-Pro	VisualPuzzles	MindCube	SITE	MathVista	MathVerse
Qwen3-VL-8B (Base)	**85.4**	**88.6**	**70.9**	69.6	55.9	47.9	41.5	41.1	77.2	62.1
CogSense-8B (Ours)	85.1	84.7	66.8	**71.4**	**58.2**	**53.6**	**42.5**	**42.1**	**79.1**	**65.7**

H Additional Qualitative Examples

More qualitative examples are shown in Figure H1. These examples demonstrate that CogSense-8B exhibits a superior capacity for high-level visual reasoning and rule induction compared to both large-scale proprietary models like Gemini 2.5 Pro and specialized baselines such as InternVL3. In complex categorization tasks, while baseline models often struggle with task comprehension or fail to distinguish between superficial visual similarities and core semantic categories, CogSense-8B accurately identifies latent action-based rules through analysis. In abstract geometric reasoning, CogSense-8B also successfully deciphers complex transformations between orthogonal and diagonal line structures, where other models fail to move beyond basic structural complexity. These results highlight the model's robustness in bridging the gap between low-level perception and multi-step cognitive reasoning, particularly in scenarios requiring precise spatial logic.

I Additional General Ability Evaluation

We further compare CogSense-8B against the base model on: (1) general vision-language benchmarks, including AI2D [58], ChartQA [95], MMStar [28], MMMU [158], MMMU-Pro [159], and VisualPuzzles [121]; (2) spatial reasoning benchmarks, including MindCube [132] and SITE [135]; and (3) mathematical reasoning benchmarks, including MathVista [89] and MathVerse [174]. As shown in Table I1, CogSense-8B remains competitive with the base model on these general-domain benchmarks, indicating that it does not overfit to the cognitive training distribution. As is commonly observed in post-training, performance fluctuations can arise across diverse downstream tasks (some may increase and some may decrease). In particular, ChartQA and MMStar rely heavily on OCR-related capabilities of VLMs and require fine-grained text extraction capabilities. This differs from our target domain of pure visual reasoning, which does not rely on textual priors embedded within images. More importantly, CogSense-8B outperforms the same-scale base model on all newly introduced spatial benchmarks, demonstrating stronger spatial generalization and more robust reasoning across evaluation settings. These results show that the architectural and training components introduced in Cognitive Supersensing yield consistent gains over the base model in the target reasoning domains while preserving broad multimodal competence.

J Discussion on LVIP Mechanism and Latent Visual Imagery Learning

LVIP Mechanism Explanation. The latent visual imagery generation is synchronized with the autoregressive next-token prediction. At each decoding step of the textual CoT, the LVIP head generates a corresponding visual latent embedding based on the current context. Therefore, the LVIP

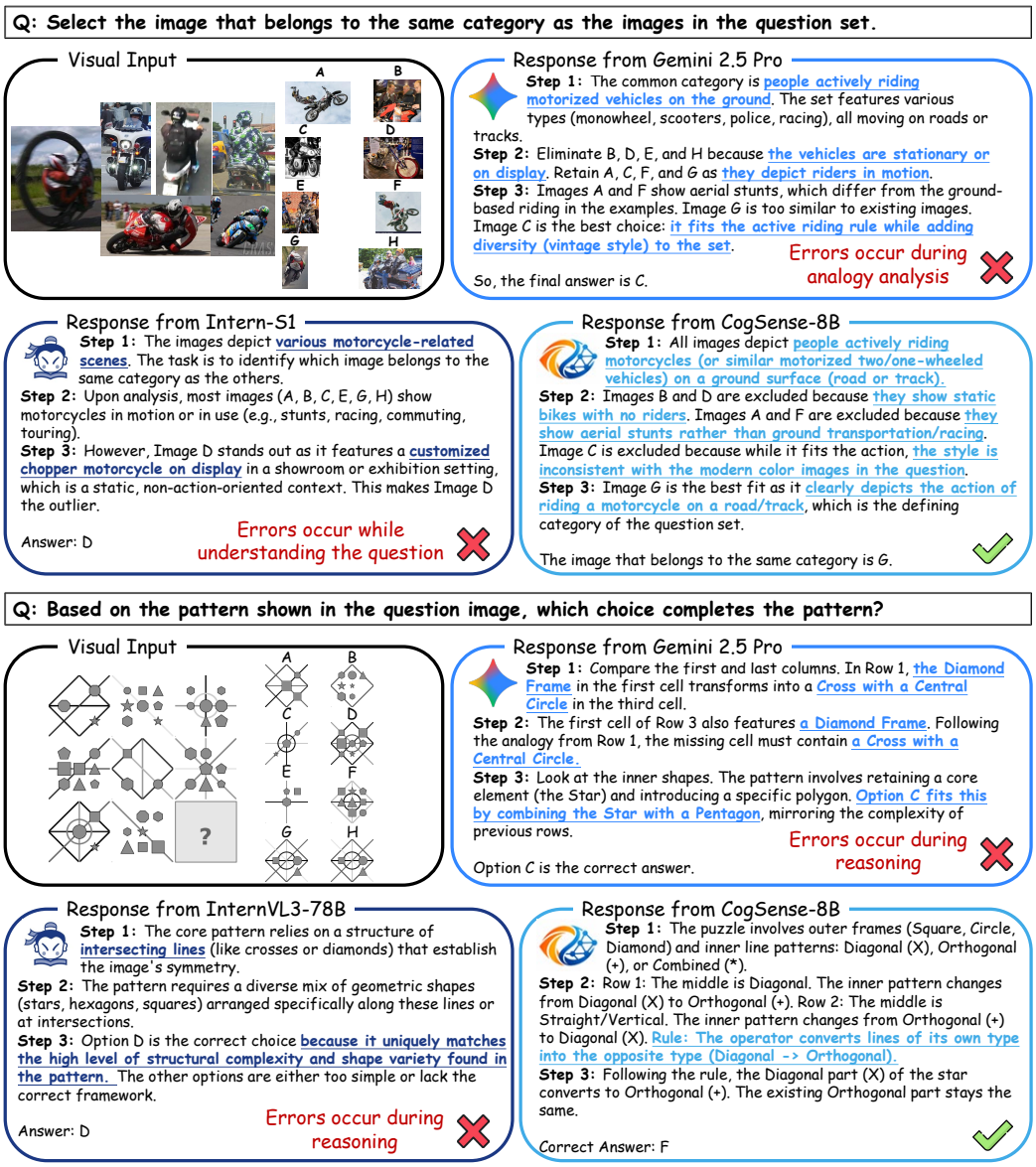

Figure H1: **Additional Qualitative Examples of Visual Cognition Reasoning Across Models.** CogSense-8B demonstrates a coherent, multi-step logical chain that closely matches the ground truth, while other models exhibit less precise or less interpretable reasoning paths.

head produces a sequence of latent visual features that changes as the reasoning progresses. The generated latent visual imagery does not merely passively follow the text's line of reasoning but actively corrects it. As detailed in Section 4, LVIP is evaluated at intermediate prefixes $[X; z_{1:t}]$, and prefixes whose latent visual states become inconsistent receive lower reward $R_{\text{lvip}}(\tau_t; X, y)$ and therefore lower probability under the learned policy $q_\theta(Z \mid X)$. In this way, LVIP influences reasoning by guiding the model toward trajectories that remain visually consistent. Through the Sub-Trajectory Balance (SubTB) [90] loss, the model is explicitly trained to assign a lower probability flow to these branching paths. In this way, LVIP trains the policy to avoid those wrong reasoning paths, and also prevents the answer-conditioned shortcut behavior through prefix-level optimization.

Analysis of Latent Visual Imagery Evolution. We track the trajectories of the intermediate latent visual imagery states during the reasoning process when the final answer is correct/incorrect. Specifically, we measure the MSE loss between the intermediate latent visual imagery $\hat{\mathbf{h}}_{y,i}$ at various steps of the CoT and the final ground-truth latent visual imagery $\mathbf{h}_y$. The results are averaged over 10 sampled trajectories. As shown in Table J1, for correct answer trajectories, the MSE loss decreases

Table J1: **Trajectories of Intermediate Latent Visual States.** We report the MSE loss between intermediate and ground-truth latent imagery across different reasoning steps. For correct answers, the latent state progressively converges, whereas it fails to converge for incorrect answers, demonstrating dynamic synchronization alongside textual reasoning.

Final Answer	Step 0	Step 10	Step 20	Step 30	Step 40
Correct Answer	0.492	0.368	0.215	0.184	0.158
Incorrect Answer	0.495	0.442	0.405	0.438	0.426

Table J2: **CoT Ablation Study Results.** We compare the performance of CogSense-8B *w/* and *w/o* intermediate CoT across CogSense-Bench. Forcing the model to generate the final answer directly in a single step (*w/o* CoT) leads to a consistent performance drop, demonstrating that step-by-step textual reasoning is crucial for guiding the latent visual imagery.

Variant	Fluid Intel.	Crystallized Intel.	Visuospatial Cog.	Mental Simu.	Visual Rout.	Avg.
CogSense-8B *w/o* CoT	61.6	87.0	66.4	65.3	45.2	70.5
CogSense-8B *w/* CoT	63.8	91.0	69.0	68.0	50.5	73.8

Table J3: **Prefix Sensitivity Ablation Study Results.** We compare the SFT-only variant and our full Stage III model (CogSense-8B) under standard inference and when forced to generate from an injected incorrect reasoning prefix. The full model exhibits substantially less performance degradation, demonstrating that Stage III training internalizes visually grounded constraints to resist incorrect intermediate textual logic.

Variant	Fluid Intel.	Crystallized Intel.	Visuospatial Cog.	Mental Simu.	Visual Rout.	Avg.
Qwen3-VL-8B SFT *w/* LVIP forced incorrect inference prefix	28.6	45.4	26.5	31.3	25.8	34.7
Qwen3-VL-8B SFT *w/* LVIP (standard inference)	55.4	88.6	61.1	61.3	44.1	68.1
CogSense-8B forced incorrect inference prefix	51.1	78.5	54.9	56.0	40.9	61.4
CogSense-8B (standard inference)	63.8	91.0	69.0	68.0	50.5	73.8

as the step increases, providing evidence that the latent visual imagery changes and converges toward the correct GT step-by-step alongside the reasoning. For incorrect answer trajectories, the MSE fails to converge to the ground-truth latent state. When the textual reasoning chain begins to hallucinate or drift into the wrong logical path, the latent visual features also reflect this internal misunderstanding.

Ablation Study on Chain-of-Thought. To validate that latent visual imagery relies on the step-by-step textual reasoning, we conduct an ablation study removing the CoT, forcing the model to generate the final answer directly in a single-step reply. As demonstrated in Table J2, removing the CoT leads to a performance drop, demonstrating that the intermediate textual reasoning is crucial to guide and refine the latent visual imagery toward the correct answer.

Ablation Study on Prefix Sensitivity. To directly test whether LVIP is merely passively following the textual chain, we perform a robustness experiment in which we inject an incorrect intermediate inference prefix and force the model to continue generation from that point. As shown in Table J3, if LVIP only acted as a passive auxiliary signal, then the model would largely inherit the erroneous reasoning and collapse similarly to the SFT-only variant. Instead, we observe a much smaller degradation for the full Stage III model compared to Qwen3-VL-8B SFT *w/* LVIP. This indicates that Stage III training with LVIP internalizes visually grounded constraints into the learned policy, making the model substantially more robust to flawed intermediate reasoning prefixes.

Ablation Study on Intermediate State Perturbation. If a model relies on an answer-conditioned shortcut, its intermediate steps should not causally drive the reasoning process. Therefore, corrupting an intermediate state should not disrupt a shortcut that directly maps to the final sequence. Following [81], we test whether each latent visual state is functionally necessary by selecting successful reasoning trajectories and injecting random noise into the latent visual state at a specific step (Step 30). The model then continues decoding from the perturbed state. We measure the MSE between the intermediate latent visual imagery $\hat{\mathbf{h}}_{y,i}$ at different CoT steps and the final ground-truth latent visual imagery $\mathbf{h}_y$, averaged over 10 sampled trajectories. As shown in Table 4a, the trajectory remains stable before the intervention. Once the Step-30 state is perturbed, downstream latent error increases, and the trajectory fails to reconverge to the clean path. This provides evidence that intermediate latent visual states play a causal role in sustaining the reasoning trajectory, rather than merely being correlated with answer confidence.

Ablation Study on Latent State Intervention. To further test whether LVIP actively steers reasoning, we conduct a latent patching experiment inspired by [82]. We intervene during intermediate decoding by swapping the latent visual state between correct and incorrect reasoning runs. We measure the change in the probability of generating the correct answer after the intervention ($\Delta p = p_{\text{after}} - p_{\text{before}}$). As shown in Table 4b, when we patch a distractor's latent state into a correct trajectory, the downstream rationale shifts toward that distractor. Conversely, patching a correct latent state into an incorrect run actively shifts the trajectory toward the correct answer. This provides strong evidence that the patched intermediate state carries vital causal information and actively steers the inference.

Ablation Study on Progressive Information Emergence. We also measure how predictive the intermediate latent visual state is of the final answer at different reasoning steps using a linear probe [1]. Specifically, we freeze CogSense-8B and train a lightweight linear classifier to predict the final correct option exclusively using the intermediate latent visual imagery $\hat{\mathbf{h}}_{y,i}$ extracted at different steps of the reasoning process. If the model were taking a trivial shortcut, the latent state would immediately identify the answer from the start. As shown in Table 4c, the accuracy increases gradually and progressively rather than appearing fully formed at early steps. This rules out the trivial shortcut hypothesis, providing evidence that the target information is genuinely constructed step-by-step via latent visual imagery as the textual logic unfolds.

Altogether, these experiments provide evidence that LVIP learns meaningful intermediate visual states, rather than functioning as an answer-conditioned shortcut or a passive byproduct of textual CoT.

K Ablation Study on the Number of Sampled Latent Rationales

To analyze the sensitivity of the hyperparameter n (the number of sampled latent rationales) in the paragraph **Bayesian Posterior over Latent Rationales** in Section 4.4 and its impact on inference latency, we conduct an ablation study by varying n during inference. As shown in Table K1, increasing the number of sampled latent rationales gradually increases the accuracy. After $n = 8$, the improvement in accuracy begins to level off and comes with a non-linear increase in inference latency. The default setting $n = 8$ provides a robust balance between predictive performance and practical deployment efficiency.

Table K1: **Rationales Quantity Ablation Study Results.** We report the average accuracy and average inference latency across different values of n. The default setting ($n = 8$) provides an optimal trade-off, securing performance gains before latency increases non-linearly.

n	Average Accuracy	Average Latency
1 (No aggregation)	71.4	1.5s
4	72.9	1.8s
8 (Default)	73.8	2.4s
16	73.9	3.8s
32	73.9	6.5s

L Ablation Study on Rationale Sampling Strategies

To ensure that the performance gains of CogSense-8B do not simply come from a larger sampling budget of rationales during inference, and to explore the effectiveness of different sampling strategies, we conduct a comprehensive ablation study comparing various sampling and selection strategies. To ensure a fair comparison, we fix the total sampling budget for all multi-sample baselines to $n = 8$. We maintain consistency across all experiments by using the same decoding temperature, maximum rationale length, and the same frozen scorer for reranking when applicable.

We evaluate the following strategies to isolate the impact of our proposed approach:

- **Greedy Decoding**: A single deterministic decoding baseline representing the standard performance without multiple samples.
- **SC-Majority**: Following the Self-Consistency method from [136], we sample $n = 8$ rationales and perform a majority vote on the final answers, treating each path with equal weight.
- **SC-Reranking**: We sample $n = 8$ rationales and use the frozen scorer to select the path with the highest reward score, using only that path's final answer.
- **Ours**: Our proposed sampling strategy described in Section 4.4.

As shown in Table L1, across all variants from the base model up to GRPO models, sampling strategies such as Self-Consistency and Scorer Reranking consistently provide marginal improvements over

Table L1: **Sample Strategies Ablation Study Results.** All $n = 8$ variants use the same number of sampled rationales. CogSense-8B remains strongest, showing that the gain is not merely due to self-consistency or reranking.

Variant	Strategy	Fluid Intel.	Crystallized Intel.	Visuospatial Cog.	Mental Simu.	Visual Rout.	Avg.
Qwen3-VL-8B (Base)	Greedy Decoding	31.2	34.8	31.0	45.3	40.9	35.5
Qwen3-VL-8B (Base)	SC-Majority	31.9	35.3	31.9	46.0	41.9	36.2
Qwen3-VL-8B (Base)	SC-Reranking	32.2	35.9	31.9	46.7	41.9	36.6
Qwen3-VL-8B SFT *w/o* LVIP	Greedy Decoding	51.1	76.6	63.7	59.3	41.9	62.3
Qwen3-VL-8B SFT *w/o* LVIP	SC-Majority	52.2	77.7	64.6	60.0	41.9	63.2
Qwen3-VL-8B SFT *w/o* LVIP	SC-Reranking	52.9	78.3	64.6	60.0	43.0	63.7
Qwen3-VL-8B SFT *w/* LVIP	Greedy Decoding	55.4	88.6	61.1	61.3	44.1	68.1
Qwen3-VL-8B SFT *w/* LVIP	SC-Majority	56.2	88.9	61.9	62.0	45.2	68.7
Qwen3-VL-8B SFT *w/* LVIP	SC-Reranking	56.5	89.1	62.8	62.0	45.2	69.0
Qwen3-VL-8B SFT *w/* LVIP + GRPO	Greedy Decoding	59.1	89.9	64.6	65.3	46.2	70.8
Qwen3-VL-8B SFT *w/* LVIP + GRPO	SC-Majority	59.4	90.2	65.5	66.0	47.3	71.3
Qwen3-VL-8B SFT *w/* LVIP + GRPO	SC-Reranking	59.8	90.2	65.5	66.0	47.3	71.4
Qwen3-VL-8B SFT *w/* LVIP + GRPO + LR	Greedy Decoding	59.8	90.2	65.5	66.7	47.3	71.5
Qwen3-VL-8B SFT *w/* LVIP + GRPO + LR	SC-Majority	60.5	90.5	66.4	67.3	48.4	72.1
Qwen3-VL-8B SFT *w/* LVIP + GRPO + LR	SC-Reranking	61.2	90.8	66.4	67.3	48.4	72.4
CogSense-8B (Ours)	Ours	**63.8**	**91.0**	**69.0**	**68.0**	**50.5**	**73.8**

Greedy Decoding. However, these improvements show diminishing returns. For instance, applying SC-Reranking to the standard SFT models cannot close the performance gap to models trained with trajectory-level optimization (GRPO). Furthermore, even when applying the SC-Reranking strategy to our strongest baseline (SFT *w/* LVIP + GRPO + LR), the average performance (72.4%) still falls short of CogSense-8B (73.8%). This firmly establishes that the superior performance of CogSense-8B is driven intrinsically by our visually grounded latent rationale optimization and targeted sampling framework, rather than merely relying on a larger sampling budget.

M Ablation Study on Teacher Models

We regenerate the full set of training rationales using Gemini 2.5 Pro [32] instead of GPT-5 [120] while keeping all other experimental settings unchanged. This yields an average accuracy of 73.5, compared to the 73.8 achieved with GPT-5-generated rationales. The near-identical performance indicates that our results are robust and not specific to the rationale style of a single teacher model, demonstrating that our method learns generalizable visual cognition patterns rather than merely adapting to distributions of specific datasets or teacher-generated rationales.

N Ablation Study on Backbone Models

To evaluate the generalizability of our proposed Cognitive Supersensing approach, we extend our experiments to incorporate more advanced vision-language foundation models. Specifically, we replace the original backbone with InternVL3.5-8B [134] and Qwen3.5-9B [104]. We retrain these new baseline models and comprehensively re-evaluate them on CogSense-Bench to isolate the impact of our method across different model configurations.

The results, detailed in Table N1, demonstrate that Cognitive Supersensing consistently and significantly improves performance across all evaluated cognitive dimensions, regardless of the underlying backbone model. Notably, applying our method to Qwen3.5-9B yields a remarkable average score of **78.3**, representing an absolute improvement of **9.2** points over the standard fine-tuned Qwen3.5-9B baseline (69.1). Similarly, CogSense augments the InternVL3.5-8B model by **11.3** points on average compared to its fine-tuned counterpart. These findings confirm that our approach is not constrained by specific architectural choices, but rather serves as a broadly effective mechanism for enhancing latent visual reasoning and multifaceted cognitive capabilities in state-of-the-art MLLMs.

O OOD Evaluation on Open-Ended VQA Benchmarks

The LVIP training objective uses candidate answer images to supervise the latent visual target and learn how to generate latent visual features. However, at inference time, LVIP does not require access to candidate answer images or any explicit multiple-choice options. Instead, it internally generates latent visual imagery that supports the model's textual reasoning, enabling the same

Table N1: **Backbone Ablation Study Results.** We evaluate the generalizability of Cognitive Supersensing across different foundation models. The results demonstrate that Cognitive Supersensing consistently yields significant performance gains regardless of the backbone architecture. Red numbers (↑.) denote the absolute improvement over the corresponding fine-tuned (FT) baseline.

Model	Fluid Intel.	Crystallized Intel.	Visuospatial Cog.	Mental Simu.	Visual Rout.	Avg.
Qwen3-VL-8B (FT)	51.1	76.6	63.7	59.3	41.9	62.3
InternVL3.5-8B (FT)	52.2	80.2	63.7	59.3	41.9	63.9
Qwen3.5-9B (FT)	56.5	89.1	61.9	62.7	46.2	69.1
CogSense (Qwen3-VL-8B)	63.8$_{\uparrow 12.7}$	91.0$_{\uparrow 14.4}$	69.0$_{\uparrow 5.3}$	68.0$_{\uparrow 8.7}$	50.5$_{\uparrow 8.6}$	73.8$_{\uparrow 11.5}$
CogSense (InternVL3.5-8B)	64.1$_{\uparrow 11.9}$	93.2$_{\uparrow 13.0}$	70.8$_{\uparrow 7.1}$	69.3$_{\uparrow 10.0}$	51.6$_{\uparrow 9.7}$	75.2$_{\uparrow 11.3}$
CogSense (Qwen3.5-9B)	69.9$_{\uparrow 13.4}$	97.3$_{\uparrow 8.2}$	68.1$_{\uparrow 6.2}$	70.0$_{\uparrow 7.3}$	53.8$_{\uparrow 7.6}$	78.3$_{\uparrow 9.2}$

Table O1: **Out-of-Domain Evaluation on Open-Ended VQA Benchmarks**. We evaluate the generalization ability on open-ended VQA tasks of CogSense-8B compared with the base model.

Model	BabyVision					MathVerse	CLEVRER
	Fine-grained Discrimination	Visual Tracking	Spatial Perception	Visual Pattern Recognition	All	w/o	Descriptive
Qwen3-VL-8B (Base)	12.7	10.8	15.0	15.0	13.1	62.1	62.4
CogSense-8B (Ours)	15.3	13.2	18.1	17.6	15.7	65.7	67.9

mechanism to be applied to both multiple-choice and open-ended tasks. To verify the open-ended generalization of CogSense-8B, we conduct an OOD evaluation on open-ended VQA benchmarks, including BabyVision [26], MathVerse [174], and CLEVRER [152]-Descriptive. Reported values represent the Pass@1 accuracy, following the settings of the original benchmarks. These evaluations contain no candidate images to select or match at inference. Thus, their gains cannot be explained by an inference-time mechanism that simply aligns the hidden state with option images. As shown in Table O1, our model consistently outperforms the baseline even on strictly open-ended VQA tasks, demonstrating that the cognitive reasoning abilities gained through image-based supervision transfer effectively to open-ended settings.

P Broader Impacts

This paper presents work whose goal is to advance the fields of machine learning and cognitive science, specifically in multimodal reasoning and visual cognition. The proposed methods aim to improve MLLMs' ability to solve complex visual reasoning tasks, which could have broad applications in education, scientific discovery, and assistive technologies. While we believe the potential benefits are substantial, we acknowledge that improved visual reasoning capabilities could be misused in contexts such as automated surveillance or manipulation. We encourage the research community to consider ethical implications and develop appropriate safeguards as these technologies advance.

Q Limitations

While CogSense-8B achieves strong visual cognitive performance at the 8B scale, our experiments focus on validating the proposed Cognitive Supersensing paradigm within a moderate-scale MLLM setting. Future work could investigate scaling properties to substantially larger 30B- or 70B-style MLLM backbones, larger training corpora, and distributed training setups.

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: The abstract and introduction accurately state the primary claims and scope of the paper, specifically: (1) proposing the Cognitive Supersensing framework and LVIP head, (2) introducing the CogSense-Bench, and (3) summarizing the experimental results demonstrating state-of-the-art performance and out-of-domain generalization. These directly map to the core contributions listed at the end of the Introduction.

   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 limitation of the absence of verification on ultra-large-scale models due to computational constraints, in a dedicated Limitations section in Appendix Q.

   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: The paper mainly focuses on a novel training paradigm, Cognitive Supersensing, and the introduction of a new benchmark, CogSense-Bench. It does not propose formal theoretical theorems, bounds, or mathematical proofs.

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: The paper provides comprehensive details to ensure reproducibility. Section 5 details the training settings, including hardware, base models, optimizer, and learning rates. Appendix C and Appendix D provide thorough documentation on the dataset composition, data extraction and reformatting pipeline, and the exact prompts used for Stage I reasoning chain generation. Finally, the abstract explicitly states our commitment to open-source the CogSense-Bench and our model weights, which can be found here.

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: We have explicitly committed in the abstract to open-sourcing the CogSense-Bench and our model weights, which can be found here. Detailed implementation settings and data curation steps are also provided in Appendix C and Appendix D to aid reproducibility in the interim.

   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: Section 5 explicitly details the experimental setup, including the baselines, evaluation metrics, base model, teacher model, hardware, and implementation specifics such as the optimizer, learning rate, and weight decay. Furthermore, detailed dataset statistics and data pipeline implementations are provided in Appendix C and Appendix D. An Ablation study on the hyperparameter (the number of sampled latent rationales) is provided in Appendix K.

   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 conduct a statistical significance analysis of the experimental results in Appendix G. We report the bootstrap confidence intervals and McNemar test results of the experiment results, validating the efficacy of Cognitive Supersensing.

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: [No]

   Justification: Section 5 specifies the hardware architecture used for our experiments (8 NVIDIA H200 GPUs). However, we do not explicitly report the detailed execution time, memory constraints, or total compute budget for each individual experimental run.

   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: The research conforms to the NeurIPS Code of Ethics. We obtained Institutional Review Board (IRB) approval for our human baseline study and ensured voluntary participation (detailed in Appendix E). Furthermore, we properly cited and respected the licenses of all utilized datasets in Appendix C and included a dedicated discussion on the broader impacts and potential ethical implications of our work in Appendix P.

   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 include a dedicated Broader Impacts section in Appendix P. This section discusses potential positive applications, such as advancements in education, scientific discovery, and assistive technologies, while also acknowledging potential negative societal impacts, including risks related to automated surveillance and manipulation.

    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: The released CogSense-8B and CogSense-Bench are specifically focused on abstract visual reasoning. The dataset is curated from established academic benchmarks containing cognitive puzzles, abstract shapes, and scientific diagrams as detailed in Appendix C, rather than uncurated internet scrapes. Consequently, these specific assets do not pose the high risks of generating realistic harmful content, privacy violations, or toxicity that would necessitate strict technical release safeguards.

    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 properly credit the creators of all existing datasets used to construct the CogSense-Bench by citing their original papers in Appendix D. We explicitly detail the specific licenses for each utilized asset in Table C1 in Appendix C.

    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: We introduce the CogSense-Bench dataset, training code, and CogSense-8B model weights as new assets. The data collection, formatting pipeline, and dataset statistics are comprehensively documented in Appendix C and Appendix D. All newly introduced code, models, and benchmark datasets are open-sourced and released.

    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: [Yes]

Justification: Appendix E details the participant recruitment strategy, the platform used (Google Forms), and explicitly states the details regarding compensation (participation was entirely voluntary and unpaid).

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: [Yes]

Justification: Section 5 and Appendix E explicitly state that the human study protocols were reviewed and approved by an Institutional Review Board (IRB) before the experiment. Because the study involved only participants answering multiple-choice visual cognitive puzzles, risks were minimal, and participation was entirely voluntary.

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: We use GPT-5 to generate training rationales in Stage I, with answer-consistency filtering and human review as described in Appendix D.2. We also use Gemini 2.5 Pro in the teacher-model ablation reported in Appendix M. The core method development in this research, the Cognitive Supersensing paradigm and LVIP head, does not involve LLMs as any important, original, or non-standard components.

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.