

Human Cognitive Benchmarks Reveal Foundational Visual Gaps in MLLMs

Jen-Tse Huang^{1*} Dasen Dai² Jen-Yuan Huang³ Youliang Yuan^{4*} Xiaoyuan Liu^{4*} Wenzuan Wang^{5*}
Wenxiang Jiao⁶ Pinjia He⁴ Zhaopeng Tu⁶ Haodong Duan⁷

Abstract

Despite significant progress on popular multimodal benchmarks, state-of-the-art Multimodal Large Language Models (MLLMs) continue to struggle with basic visual reasoning tasks that are trivially solved by humans, such as recognizing spatial relationships. To systematically investigate this gap, we introduce **VisFactor**, a benchmark that digitizes 20 vision-centric subtests from a well-established cognitive psychology assessment. These subtests span four core domains of human visual cognition: (1) Visualization and Spatial Processing, (2) Perceptual and Closure, (3) Memory, and (4) Reasoning. We evaluate 20 frontier MLLMs from GPT, Gemini, Claude, LLaMA, Qwen, and SEED families. The best-performing model achieves a score of only 25.19 out of 100, with consistent failures on tasks such as mental rotation, spatial relation inference, and figure-ground discrimination—regardless of model size or prompting strategy. These findings suggest that current MLLM performance gains on high-level benchmarks do not reflect human-like low-level visual cognition, challenging the assumption that large-scale pretraining naturally induces gestalt-like perceptual capabilities. The dataset and evaluation toolkit are publicly available at: <https://github.com/CUHK-ARISE/VisFactor>.

1. Introduction

Multimodal Large Language Models (MLLMs) have rapidly advanced the state of multimodal artificial intelligence, delivering impressive results in text recognition (Liu et al., 2024b; Chen et al., 2025), mathematical reasoning (Yang

*Work done when interning at Tencent AI Lab ¹Johns Hopkins University ²Chinese University of Hong Kong ³Peking University ⁴Chinese University of Hong Kong, Shenzhen ⁵Renmin University of China ⁶Tencent ⁷Shanghai AI Lab. Correspondence to: Wenzuan Wang <jwxwang@gmail.com>, Haodong Duan <duanhaodong@pjlab.org.cn>.



Figure 1. **VisFactor** integrates 20 subtests adapted from standardized human cognitive assessments. Subtests are organized into four major domains and weighted by test case count (shown numerically), which determines each segment’ visual area.

et al., 2024; Peng et al., 2024), and even clinical decision support (Azad et al., 2023; Buckley et al., 2023). On holistic leaderboards such as MM-Bench (Liu et al., 2024a), frontier models like Gemini-2.5-Pro (Kavukcuoglu, 2025) now surpass 88.9 out of 100, fueling optimism that large-scale pre-training may already endow machines with near-human visual cognition.

Closer inspection paints a different picture. Targeted studies reveal that MLLMs still falter on visual reasoning tasks that human novices solve effortlessly (Fu et al., 2024). Ramakrishnan et al. (2025) reports near-random accuracy on mental rotation test and maze completion test. Why do models that *see* so well in benchmarks apparently fail to *perceive*? A key limitation lies in today’s evaluation culture: most benchmarks emphasize downstream task performance and aggregate scores, but seldom probe the *foundational* visual faculties that underlie human reasoning.

Human vision develops hierarchically: low-level perceptual skills—figure–ground segregation, object permanence, spatial scanning—serve as scaffolds for higher-order reasoning. Cognitive psychology therefore decomposes vision into latent factors that can be measured independently. The *Factor-Referenced Cognitive Test* (FRCT) battery (Ekstrom & Harman, 1976) operationalizes this idea, mapping psychometric factors to narrowly defined subtests. In contrast to omnibus IQ scales, the FRCT delivers a fine-grained cognitive profile, making it ideal for diagnosing which capacities an MLLM truly possesses.

We introduce **VISFACTOR**, the first benchmark that ports 20 vision-centric FRCT subtests to an automated, image–text setting, spanning four cognitive domains: (1) visualization and spatial processing, (2) perceptual and closure, (3) memory, and (4) reasoning (Fig. 1). Prior multimodal benchmarks (Ramakrishnan et al., 2025; Fu et al., 2024) often rely on four-option multiple-choice (25% chance) or binary yes/no (50% chance) formats, enabling models to reach non-trivial scores through random guessing or option-position biases. To preclude such shortcuts, we generate at least four rule-based variants for every **VISFACTOR** item and deliberately diversify the correct-answer distribution (*e.g.*, multiple-choice keys are not always “A,” yes/no items are not disproportionately “Yes”). This design lowers the overall chance-level accuracy to 2.9/100, ensuring that any success on **VISFACTOR** reflects genuine reasoning rather than lucky guesses.

We evaluate 20 frontier models drawn from GPT (Hurst et al., 2024; OpenAI, 2025a), o-series (OpenAI, 2025b), Gemini (Kavukcuoglu, 2025), Claude (Anthropic, 2025b), LLaMA (Meta, 2024), Qwen (Bai et al., 2025), and SEED (ByteDance, 2025) families. Despite advanced prompting strategies such as Chain-of-Thought (CoT) (Kojima et al., 2022; Wei et al., 2022), the best model scores **25.19** out of 100. Systematic failures emerge on mental-rotation, spatial-relation, and figure–ground tasks, independent of model size or architecture. The original FRCT contains a finite set of items, raising the risk that future models may overfit by training directly on the public benchmark. To future-proof **VISFACTOR**, we focus on the subtests where current models perform poorest and implement a generator that produces unlimited, difficulty-controlled instances in the style of the FRCT. Item parameters (*e.g.*, rotation angle, occlusion level, distractor similarity) can be modulated to create graduated test suites, enabling longitudinal tracking without saturating the benchmark.

Our contributions are as follows:¹

¹**VISFACTOR** is implemented with VLMEvalKit (Duan et al., 2024) and available at <https://github.com/ CUHK-ARISE/VISFactor>.

- **Factor-grounded evaluation.** We present the first benchmark that grounds MLLM assessment directly to human cognitive factors, bringing psychometric rigor to multi-modal evaluation.
- **Complete, future-proof framework.** We digitize every FRCT vision item, devise rule-based variant generation with balanced answer keys, and introduce controllable-difficulty item synthesis for the hardest subtests.
- **Comprehensive landscape study.** We benchmark twenty state-of-the-art MLLMs, offering a panoramic view of current capabilities and pinpointing cognitive gaps that chart a roadmap for future research.

2. VISFACTOR Design and Implementation

This section introduces how we select tests from FRCT (§2.1), how to fit the tests to MLLMs (§2.2–§2.3), and how we generate more difficulty-controllable test cases (§2.4).

2.1. Test Selection and Justification

The original FRCT battery comprises 72 subtests. We exclude those that cannot be assessed with a vision–language interface whose output is text only: (1) **Image-production tasks** (4 subtests): Figural Fluency (FF1–FF3) and Spatial Scanning (SS1) ask participants to draw or trace; this is incompatible with text-only output. (2) **Speech-dependent tasks** (3 subtests): Memory Span (MS1–MS3) require subjects to write down what they hear and therefore probe speech-to-text rather than visual cognition. In the remaining 65 subtests, 45 of them can be completed with pure text input. Those demanding visual reasoning but accept text answers form our benchmark, **VISFACTOR**. The 20 subtests cover 10 FRCT factors: Closure Flexibility (CF), Closure Speed (CS), Induction (I), Associative Memory (MA), Visual Memory (MV), Perceptual Speed (P), Logical Reasoning (RL), Spatial Orientation (S), Spatial Scanning (SS), and Visualization (VZ). Figure 2 shows example questions and answers of each subtest.

2.2. Digitization and Prompt Design

Instructions. Directly feeding the human-oriented FRCT instructions to MLLMs prove verbose and occasionally ambiguous. We therefore ask GPT-4o and Gemini-2.5-Flash (via web applications) to summarize each instruction set to its minimal, MLLM-friendly form. A human annotator reconcile the two summaries with the originals, producing a concise final prompt for every subtest.

Questions and Answers. All images are captured at 300 dpi and cropped to the region containing only the task stimuli (no additional texts). Ground-truth answers are extracted verbatim from the FRCT manuals.

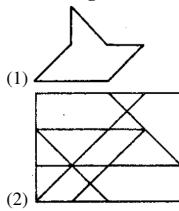
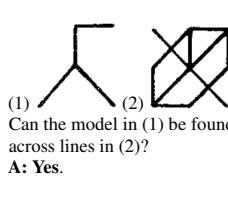
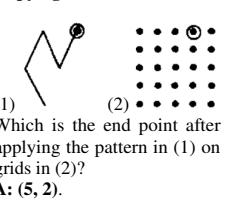
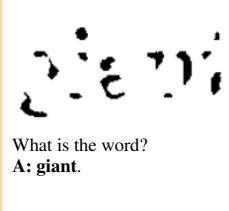
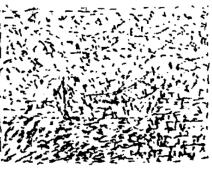
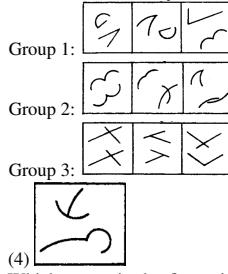
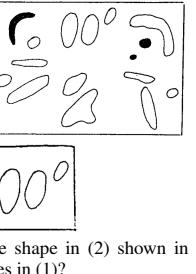
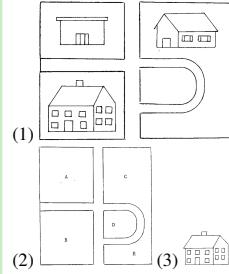
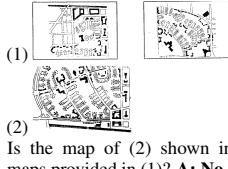
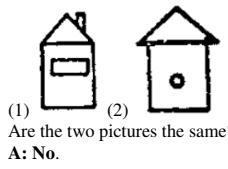
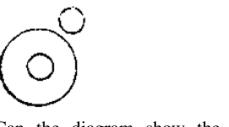
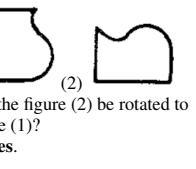
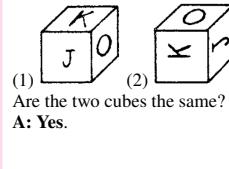
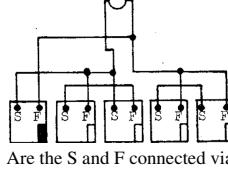
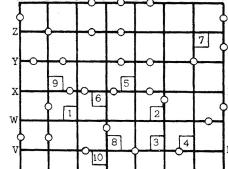
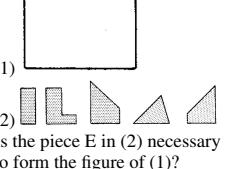
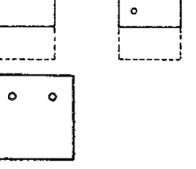
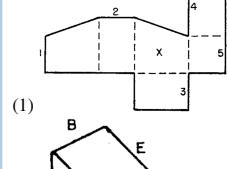
CF1 Hidden Figures Test  (1) Can the figure in (1) be found in the pattern in (2)? A: No. (2)	CF2 Hidden Patterns Test  (1) Can the model in (1) be found across lines in (2)? A: Yes. (2)	CF3 Copying Test  (1) Which is the end point after applying the pattern in (1) on grids in (2)? A: (5, 2). (2)	CS1 Gestalt Completion Test  What is the object? A: A hammer head.	CS2 Concealed Words Test  What is the word? A: giant.
CS3 Snowy Pictures  What is the object? A: A boat.	I3 Figure Classification  Group 1: Group 2: Group 3: (4) Which group is the figure in (4) from? A: 2.	MA1 Picture-Number Test  (1) (2) Which number is the object in (2) related to in (1)? A: 73.	MV1 Shape Memory Test  (1) (2) Is the shape in (2) shown in shapes in (1)? A: Yes.	MV2 Building Memory  (1) (2) (3) Should the building in (3) be in place E of map (2) according to map (1)? A: No.
MV3 Map Memory  (1) (2) Is the map of (2) shown in maps provided in (1)? A: No.	P3 Identical Pictures Test  (1) (2) Are the two pictures the same? A: No.	RL2 Diagramming Relationships  Can the diagram show the relationship between: Desks, furniture, pencils? A: Yes.	S1 Card Rotations Test  (1) (2) Can the figure (2) be rotated to figure (1)? A: Yes.	S2 Cube Comparisons Test  (1) (2) Are the two cubes the same? A: Yes.
SS2 Choosing A Path  Are the S and F connected via the white circle in box E? A: No.	SS3 Map Planning Test  Which building does the shortest route between F to T pass? A: 5.	VZ1 Form Board Test  Is the piece E in (2) necessary to form the figure of (1)? A: Yes..	VZ2 Paper Folding Test  Does the figure in (2) correctly show the paper after folding and punching via steps in (1)? A: No.	VZ3 Surface Development Test  Which edge in the 3-D figure is related to edge 5 in the 2-D figure after folding? A: H.

Figure 2. VISFACTOR comprises 20 vision-centric cognitive subtests. Each task is designed to isolate core factors of human visual cognition, covering 10 distinct factors in total. The subtests are converted into either yes/no questions or fill-in-the-blank questions according to §2.3. Example stimuli, questions, and ground-truth answers are shown for each task.

2.3. Reducing Chance-Level Accuracy

To prevent inflated scores from lucky guesses, we modify test formats as follows, except CF3 (25-way), MA1 (21-way) and all fill-in-the-blank subtests (CS1-CS3) that already exhibit $\leq 5\%$ random success. The average random guessing performance is reduced from 22.47% to 2.89%, with no single test exceeding 6.25%.

1. **Decomposed multiple choice:** For seven subtests with five options (CF1, MV2, P3, RL2, SS2, VZ1, VZ2), we pose *one yes/no query per option* and require the model to answer *all* correctly for credit. Chance accuracy thus drops from 25% to $(0.5)^5 \approx 3.13\%$.
2. **Grouped-consistency items:** Three subtests repeatedly probe the same latent feature across a small cluster of items. We aggregate each cluster and award credit only if *all* constituent items are correct. This applies to: (i) CF2 Hidden Patterns Test—400 binary items grouped into 80 sets of five; chance $(0.5)^5 \approx 3.13\%$. (ii) I3 Figure Classification—eight figures to be classified into two or three groups; chance $\approx 0.23\%$. (iii) S1 Card Rotation Test—eight judgments of the same card; chance $(0.5)^8 \approx 0.39\%$.
3. **Symmetry variants:** MV1, MV3 and S2 originally ask whether figure A matches figure B. We generate three variants per item—“A differs from B”, “B matches A”, “B differs from A”—so that “yes” and “no” answers are balanced, preventing easy success by models that consistently answer yes or no. The probability of guessing all three correctly by chance is $(0.5)^4 = 6.25\%$.
4. **Specialized rewrites:** (i) SS3 (Map Planning Test). Each item asks participants to find the building number that the shortest path between a *start* and an *end* point passes in a map. Exchanging start and end leaves the correct answer unchanged. We therefore require the model to answer *both* directions correctly, lowering chance from 10% to 1%. (ii) VZ3 (Surface Development Test). Each item asks: which 3-D edge corresponds to the marked 2-D edge after folding? Since multiple 2-D edges may map to the same 3-D edge, simply swapping the query direction (asking which 2-D edge matches a given 3-D edge) would introduce one-to-many ambiguity and ill-defined ground truth. Therefore, we add additional questions asking whether a pair of 2-D and 3-D edges are the same, resulting in all “yes” ground truth. To create “no” pairs, we generate questions with cyclic-permuted 3-D edge labels (*e.g.*, $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E \rightarrow A$). MLLMs receive credit only if they correctly answer the fill-in-the-blank question and both yes/no questions; chance $14.6/4 = 3.65\%$.

2.4. Synthetic Augmentation

A subset of tests—CF1–CF3, CS1–CS3, MA1, S1–S2, SS3, VZ1–VZ2—admits parametric generation.

CF1: Hidden Figures Test. We model each pattern as a graph $G = (V, E)$ embedded on an axis-aligned $m \times n$ lattice whose admissible edges join adjacent vertices (4-neighbour plus the two diagonals). Generation starts by deterministically adding the perimeter edges, thereby fixing a closed bounding rectangle and seeding a single connected component. The target edge count is then drawn from $k \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = \rho|E|$ and $\sigma = \rho_{\text{std}}|E|$ for user-specified density $\rho \in (0, 1]$ and ρ_{std} , and clipped to $[0, 1] \cdot |E|$. For sub-pattern detection we represent the user-supplied “model” as its own edge set and enumerate all translations obtained by aligning any model vertex with any pattern vertex; containment reduces to a constant-time subset test per translation, which is tractable for the small grids used here and yields exact, translation-invariant matches without recursion or graph isomorphism search.

CF2: Hidden Patterns Test. We introduce a graph-based generator that operates on an $m \times n$ lattice. We first enumerate the complete set \mathcal{E} of admissible edges—unit horizontal, vertical, and diagonal connections between adjacent lattice nodes—yielding $E = |\mathcal{E}|$ potential segments. To guarantee global connectivity, we draw a uniformly random spanning tree $T \subset \mathcal{E}$ by performing a depth-first search with randomized successor order; this yields exactly $N - 1$ edges, where $N = mn$ is the number of nodes. Desired edge density is controlled by sampling a target count $k \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = \rho E$ and $\sigma = \rho_{\text{std}} E$ for user-specified density $\rho \in (0, 1]$ and ρ_{std} ; the sample is clipped to $[N - 1, E]$. We then augment T with $k - (N - 1)$ additional edges drawn without replacement from $\mathcal{E} \setminus T$, producing a connected graph $G = (V, E_G)$ whose expected density equals ρ .

CF3: Copying Test. We develop a procedural grid-walk generator that produces paired images. Each instance begins by laying out an $m \times n$ lattice whose node coordinates are computed analytically from a single size parameter, ensuring scale-invariance across resolutions. A start node is selected uniformly at random and a self-avoiding walk is grown whose length is drawn from a user-specified interval $[\text{min_steps}, \text{max_steps}]$. At every extension step, the candidate set comprises all yet-unvisited lattice nodes; candidates that would yield a line segment collinear with any existing segment in the path are deterministically excluded via a zero-cross-product test, preventing visual overlap and ensuring topological diversity. Two images are rendered, a reference grid with the start node circled, and a path image of identical dimensions that shows only the start node and the resulting non-collinear walk.

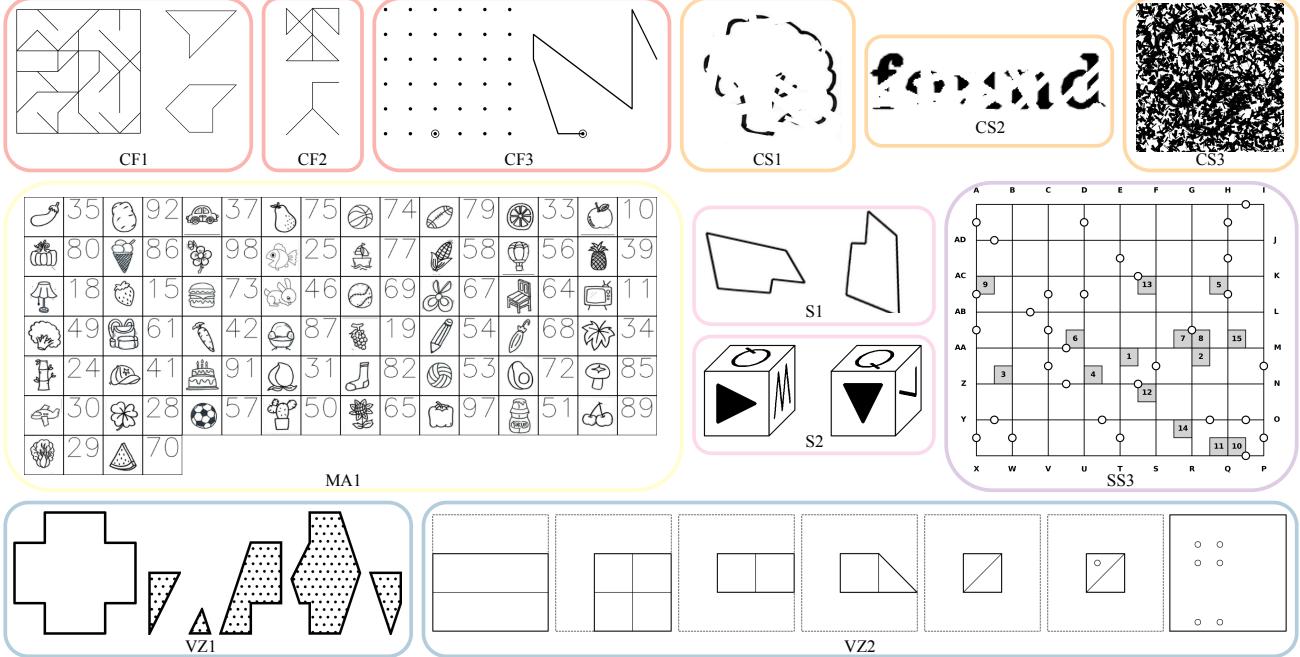


Figure 3. Samples of our generated images. We can dynamically adjust test difficulties in **ViSFACTOR**. For example, the grid size of CF3 is changed to 6×6 instead of 5×5 , and 8×9 instead of 7×8 for SS3.

CS1: Gestalt Completion Test. We begin by curating object silhouettes and their labels from public image repositories. Each image is partially occluded with randomly oriented white strokes whose number and width scale linearly with a severity coefficient $s \in [0, 1]$.

CS2: Concealed Words Test. We synthesize a tunable corpus of occluded word images by sampling from the `top_nlist` in the `wordfreq` Python library, retaining alphabetic tokens whose lengths fall within a user-defined interval and converting them to lower-case. Each word is rendered on a white canvas and then obfuscated by superimposing straight white line segments and circular blotches drawn at random positions. The number, thickness, and radius of these artifacts increase linearly with a continuous severity parameter $s \in [0, 1]$, providing precise control over the level of visual concealment.

CS3: Snowy Pictures. Building on the silhouettes and labels introduced in CS1, we corrupt every input image in two successive steps. First, we overlay n_r white rectangles whose side lengths are sampled uniformly up to a fixed fraction of the image’s shorter edge, disrupting local continuity. Next, we draw n_ℓ short, randomly oriented black line segments that imitate dense, edge-like clutter. Both n_r and n_ℓ scale linearly with a severity parameter $s \in [0, 1]$.

MA1: Picture-Number Test. Also building on the source from CS1, we first draw N unique items without replace-

ment and an equal-sized set of distinct two-digit integers $\{10, \dots, 99\}$. The two cells are concatenated horizontally to form an atomic pair, and all pairs are then tiled row-major into an $r \times c$ grid with $rc \geq N$ and $|r - c|$ minimized to approximate isotropy, yielding a visually balanced layout regardless of N . A uniformly random pair is sampled to provide a query image and its label, while the full canvas supplies rich contextual clutter.

S1: Card Rotations Test. We devise a lightweight generator that first samples a simple, non-self-intersecting polygon by drawing *i.i.d.* polar radii and sorted angles, and repeatedly rejecting candidates whose (i) shortest edge falls below a minimum-length threshold and (ii) consecutive edge-length differences are within a tight tolerance—two filters that jointly suppress near-symmetries and visually imperceptible edges. We optionally apply a horizontal mirror, then rotate it by a uniformly random angle before centrally cropping back to the original spatial extent. From every base polygon we generate N views and record a binary label indicating whether the transformation involved only rotation (true) or a mirror-plus-rotation (false).

S2: Cube Comparisons Test. To decide whether two partial observations correspond to the same physical cube, we cast the problem as a constrained search over the 24 right-handed orientations of a cube in \mathbb{Z}^3 . We first “pin” the first view as the reference orientation—its Up, Front and Right faces become the intrinsic Up, Front, Right faces of

the cube—which lets us record its three symbols and their rotations in a baseline face–rotation table. For each of the 24 global orientations we then (i) map the observer’s local axes to intrinsic cube faces via simple cross-product geometry, (ii) transform the second view’s reported rotations into the reference frame by adding a pre-computed 90° offset that aligns local “Up” vectors, and (iii) enforce two consistency constraints: (a) the same intrinsic face observed twice must carry identical symbols whose rotations are equivalent under the symbol’s symmetry class (4-fold, 2-fold, or asymmetric), and (b) a symbol may not appear on two different faces. Finally, we randomly generate such three-face views and render them as perspective-correct 3-D cube images.

SS3: Map Planning Test. We model the city layout as a rectangular $m \times n$ lattice in an undirected graph, where each vertex represents a street intersection and each edge a unit-length street segment. From the fully connected lattice we remove a user-specified fraction r of edges, chosen uniformly at random, and tag their mid-points as circular “road-blocks,” thereby enforcing non-traversable segments while preserving the geometry for visualization. N_B quarter-square buildings are sampled without replacement from the $(m - 1)(n - 1)$ grid cells, along with the two edges each of them touches. Perimeter intersections are labeled in clockwise order using spreadsheet-style indices (A–Z, AA, AB, . . .), after which start–end terminals are selected by random permutation until exactly one shortest path exists between them, which guarantees uniqueness while avoiding exhaustive search. The final instance thus comprises a sparse planar graph with a provably unique geodesic, alongside metadata for blocked edges, buildings and perimeter labels.

VZ1: Form Board Test. We design an automatic pipeline that transforms an arbitrary lattice-defined polygon into a “dissect-and-assemble” puzzle while guaranteeing a unique solution under rotation and translation. The target shape is first specified on an $n \times n$ integer grid as an ordered list of boundary edges. A random integer $k \in \{3, 4, 5\}$ determines the number of genuine solution pieces. Starting from the full polygon, we iteratively bisect the currently largest fragment with straight grid-aligned cuts whose slopes are limited to $+\infty, 0, \pm 1, \pm 2, \pm 3$. Each cut is accepted only if it produces two valid polygons, and the process terminates as soon as k fragments are obtained. To generate the remaining $5 - k$ distractor pieces, we re-cut one randomly chosen solution fragment, rejecting candidate fragments whose areas coincide with any existing piece, thereby ensuring that no spurious subset of distractors can reconstruct the target.

VZ2: Paper Folding Test. Starting from a unit-square sheet discretized into an $n \times n$ grid, our algorithm iteratively selects a random fold axis—horizontal, vertical, or an arbitrary offset diagonal of the form $y = \pm x + c$. At each

step, the square is partitioned by this axis; the half-plane judged closest to the sheet’s geometric center remains stationary, while the opposite half is reflected via an analytic mapping that preserves affine structure. Crucially, we maintain (i) a “Polygon” describing the current outer outline, (ii) an ordered list of internal edges and crease lines, and (iii) the exact set of point holes. These entities are updated by reflecting only those primitives that lie on the moving half and clipping fold-axis segments to the unfolded outline, guaranteeing topological correctness even for degenerate or off-center folds. The complete state history enables deterministic reverse unfolding to generate the answer: holes are “back-propagated” by conditional reflection.

3. Experiments

3.1. Settings

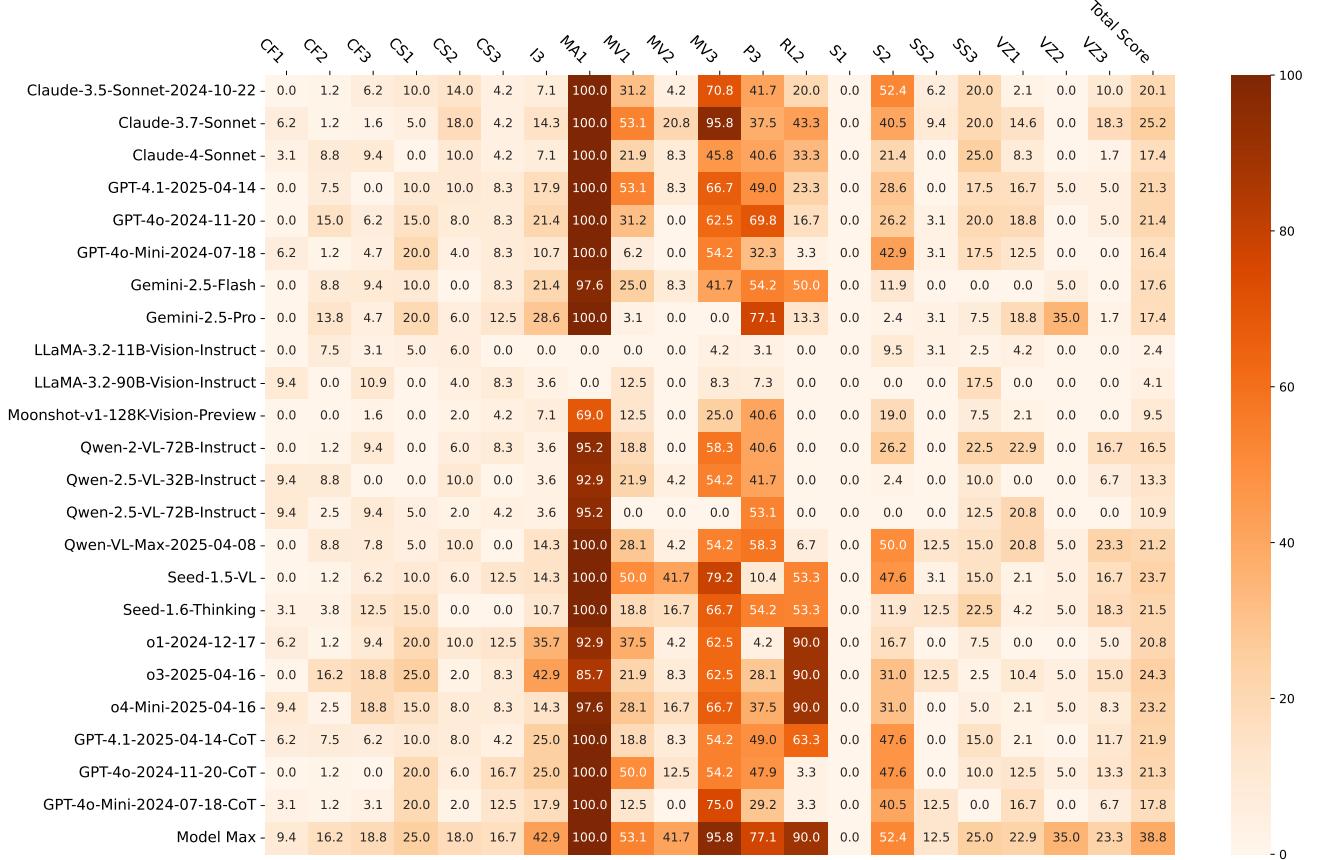
Models. We evaluate 20 models: GPT-4o (Hurst et al., 2024), GPT-4o-Mini (OpenAI, 2024), GPT-4.1 (OpenAI, 2025a), Gemini-2.5-(Pro, Flash) (Kavukcuoglu, 2025), Claude-Sonnet-(3.5 (Anthropic, 2024), 3.7 (Anthropic, 2025a), 4 (Anthropic, 2025b)), Qwen-2-VL (Wang et al., 2024a), Qwen-2.5-VL-(32B, 72B) (Bai et al., 2025), Qwen-VL-Max (Team, 2024), Seed-1.5-VL (Guo et al., 2025), Seed-1.6 (ByteDance, 2025), Moonshot-V1-128K-Vision-Preview (MoonshotAI, 2025), LLaMA-3.2-Vision-(11B, 90B) (Meta, 2024), o1 (Jaech et al., 2024), o3 (OpenAI, 2025b), and o4-Mini (OpenAI, 2025b).

Hyper-parameters and Prompts. We set the temperature to 0 for all models, except Qwen (minimum temperature 0.01) and LLaMA-3.2 (temperature 0.6). For Qwen, Top-P is set to 0.001; for LLaMA-3.2, Top-P is set to 0.9. The thinking budget is configured as *high* for Gemini-2.5 and o-series models. Greedy decoding is used as the default sampling strategy. All models are accessed via their official APIs, except LLaMA-3.2, which is run locally. In our implementation, the retry count is set to 3, allowing each case up to three retries before being marked as a failure. All test cases are conducted in a zero-shot setting. The exact prompts are provided in §A of the appendix.

3.2. Results on Original Tests

Most existing models perform poorly on the VISFAC-TOR benchmark. Among the 20 evaluated frontier models, *Claude-3.7-Sonnet* achieves the highest overall score, but only reaches 25.2 out of 100. Even when aggregating the best-performing models across individual subtests, the combined score is just 38.8. Models generally perform well on memorization tasks (MA1, MV1–MV3), indicating a strong ability to attend to relevant context in the input. A breakdown of top-performing models by subtest reveals distinct strengths: (i) OpenAI’s o-series models excel at reasoning

Table 1. The performance of 20 models on **VISFACTOR**. The bottom row shows the highest scores achieved by any model, while the rightmost column shows the total score. Darker scores show higher scores. The best model is Claude-3.7-Sonnet.



tasks (I3, RL2). They also perform best on CF1–CF3 and CS1, demonstrating superior recognition of lines and edges. (ii) Google’s Gemini leads on P3 and VZ2, particularly excelling at VZ2, which requires precise spatial localization to identify holes in paper. (iii) Qwen leads on SS2, VZ1, and VZ3, indicating strong mental imagery capabilities for shape splicing and folding. (iv) Claude performs best on CS2, MV1, MV3, S2, and SS3. (v) Seed achieves the top score on CS3 and MV2.

Model size and recency do not guarantee superior performance. For example, Qwen-2.5-32B outperforms Qwen-2.5-72B, and Qwen-2-72B also surpasses Qwen-2.5-72B. Similarly, Claude-3.7 outperforms Claude-4, and Seed-1.5 exceeds Seed-1.6. While there are exceptions—such as GPT-4o outperforming GPT-4o-Mini, and o3 surpassing o1—performance on **VISFACTOR** shows no consistent correlation with model scale or version. These results suggest that core visual capabilities may be underemphasized in current model development pipelines.

CoT offers limited benefits. We evaluate the effect of CoT prompting across three GPT models. While CoT provides some improvements, the gains in overall performance are

marginal. This aligns with recent findings showing that CoT does not universally enhance model performance; in fact, certain cognitive tasks may exhibit degraded performance with CoT (Liu et al., 2025a). Specifically, we observe declines in performance on perceptual and closure tasks (P3, CS2) and spatial visualization tasks (SS3, VZ1). Conversely, CoT consistently improves performance on reasoning tasks such as I3 and RL2, consistent with prior results (Sprague et al., 2025).

The “Middle Score Anomaly” (Babaiee et al., 2025) is also observed in our **VISFACTOR.** This phenomenon refers to models unexpectedly achieving intermediate performance—neither random nor near-perfect—on tasks that are extremely easy for humans. For instance, the Identical Pictures Test (P3) simply requires determining whether two images depict the same object. Humans can either solve this task almost perfectly or fail entirely (*i.e.*, perform at chance level if they lack the necessary perceptual ability). It would be highly unusual for a human to achieve, say, 70% accuracy on this task—suggesting partial understanding but inexplicable failures. However, we observe that most models obtain 30–50% accuracy on P3, while random guessing

Table 2. The performance of the GPT-4.1 model on the generated subsets in **VISFACTOR**. The “Original” row reports performance on the original FRCT questions. The “Normal” row uses the same configuration as the original questions. The “Easy” and “Hard” rows correspond to questions that are modified to be easier and more difficult, respectively.

	CF1	CF2	CF3	CS1	CS2	CS3	MA1	SS1	SS2	VZ1	VZ2	Total Score	
Easy -	3.1	13.8	15.6	40.0	78.0	45.8	88.1	0.0	2.4	45.0	14.6	0.0	28.9
Hard -	0.0	17.5	4.7	35.0	52.0	25.0	78.6	0.0	0.0	32.5	18.8	0.0	22.0
Normal -	0.0	12.5	4.7	35.0	76.0	16.7	90.5	0.0	0.0	30.0	12.5	0.0	23.2
Original -	0.0	7.5	0.0	10.0	10.0	8.3	100.0	0.0	28.6	17.5	16.7	5.0	21.3

yields only 3.13%. We interpret this as evidence that current models lack genuine reasoning capabilities, at least in the context of the tasks presented in **VISFACTOR**.

3.3. Results on Generated Tests

Using our generation algorithms, we first construct a “Normal” subset in which each configuration closely mirrors the original FRCT questions. We then create “Easy” and “Hard” subsets by systematically adjusting parameters that modulate task difficulty. For instance, we vary the grid size for CF1, CF2, CF3, SS3, and VZ1; the noise severity for CS1, CS2, and CS3; the number of item pairs to be memorized in MA1; and the number of folds in VZ2.

We evaluate the GPT-4.1-2025-04-14 model, and the results are presented in Table 2. The model’s performance increases progressively across the easy, normal, and hard subsets. Our key findings are as follows: (1) CS1–3 (object and word recognition under noise): The model achieves higher accuracy on our generated datasets compared to the original ones. We attribute this to our selection of commonly encountered objects in daily life, which likely reduces recognition difficulty. Moreover, our framework supports dynamic image updates, allowing the tests to be refreshed as needed in the future. (2) MA1 (memory test): The original version requires memorizing 21 image-number pairs, a task on which the model achieves 100% accuracy. In contrast, our hard version increases the number of pairs to 50, resulting in a substantial performance drop, highlighting the increased challenge. (3) VZ2 (paper folding test): The original dataset includes questions based on one to three folds. Our version expands this to include up to five folds, significantly increasing task complexity. The model fails to answer any of these questions correctly. These results demonstrate that our generated dataset effectively supports dynamic adjustment of test difficulty, making it suitable for evaluating increasingly capable models.

4. Failure Analysis

4.1. Visual Comparison Or Concept Recognition?

How Models Master MA1? Given that models achieve very high accuracy on this memory test, we investigate

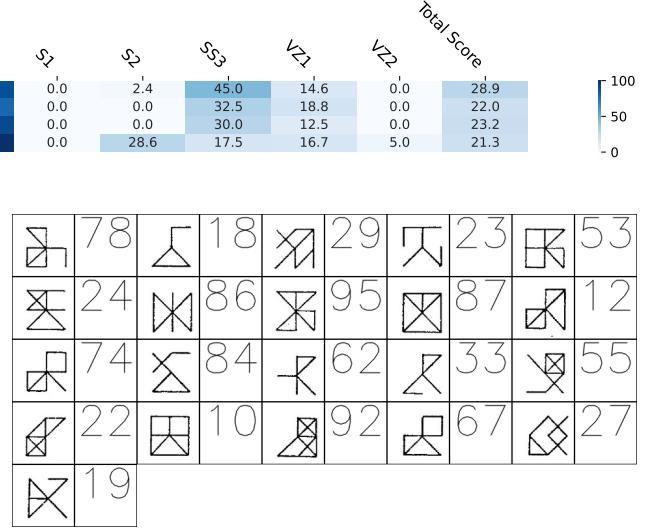


Figure 4. An example of our generated MA1 using CF2 figures.

how they are able to solve these problems. An intuitive hypothesis is that models translate visual cues into high-level, human-interpretable concepts (*e.g.*, “soccer,” “chair,” “fish”) and memorize the concept–number pairs, rather than the raw image patterns. To test this hypothesis, we use CF2-generated images, which consist only of lines arranged in a 3×3 grid, to create MA1 test cases via our automatic generation algorithm (see Fig. 4 for an example). When evaluated on these inputs, GPT-4.1’s accuracy drops sharply—from 90% to 23%—indicating a strong reliance on interpretable concepts. To ensure that the performance drop is not simply due to distributional shift, we generate extreme yet valid visual combinations using diffusion models (*e.g.*, “a horse on the moon”). In these cases, the model maintains high accuracy, further supporting our hypothesis: the model performs well as long as the visual input can be mapped to familiar, conceptual categories. These results also suggest that models struggle to interpret abstract visual patterns such as the line-based CF2 stimuli, reinforcing the idea that their success depends on concept recognition rather than low-level perception.

This hypothesis is further supported by the analysis of the P3 task, which reveals that high-performing examples typically involve easily verbalizable content, whereas failures are associated with visually complex and linguistically demanding patterns. These findings suggest that apparent visual comparison abilities may primarily reflect advanced textual reasoning applied to visual descriptions, rather than authentic visual processing.

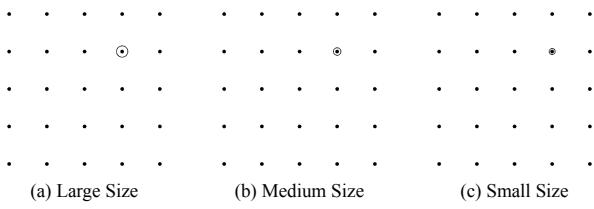


Figure 5. Our generated CF3 figures with different marker sizes.

4.2. Visual Recognition: A Key Bottleneck

Rely on Accurate Textual Descriptions. Our comprehensive evaluation reveals a stark contrast between models’ strong textual reasoning capabilities and their significantly weaker visual perception performance. This disparity is exemplified by the CF3 task: when models are provided with textual descriptions of line segments (defined by starting coordinates and displacement vectors), GPT-4.1 achieves perfect accuracy (100%). In contrast, performance drops sharply when the same information has to be inferred from visual inputs, with accuracy falling to just 6.2%—and no model exceeding 18.8%.

Fail to Recognize Visual Details. In the SS2 task, models consistently fail to distinguish between intersecting lines with explicit junction markers versus those without visual indicators. More critically, our automated generated CF3 test cases (illustrated in 5) reveal that start-point identification accuracy deteriorates systematically with marker size variation: from 92% with large circular markers to 80% with medium markers, and ultimately 68% with small markers. This degradation pattern suggests fundamental constraints in the models’ visual attention mechanisms, where reduced visual saliency directly compromises recognition performance.

Additionally, models struggle to focus effectively on key regions, resulting in missed information. For example, in CS2, the task involves identifying the partially erased word “women.” Correct identification of the first character requires recognizing the faint stroke in the lower left corner that differentiates “w” from “v.” Similarly, identifying the fifth character as “n” relies on detecting a small vertical stroke in the lower right corner of the letter. Models, however, misclassify these characters as “v” and “r,” respectively, indicating its limited ability to prioritize critical local features.

4.3. Low Sensitivity to Length, Angle, and Scale

Models exhibit notable limitations in processing geometric shapes, particularly in assessing length and proportion. In the CF3 (Copying Test), the task is to replicate lines from the left side onto a 5×5 dot matrix on the right. While

models can approximate line directions, they frequently err in determining their lengths. Similarly, in the VZ1 (Form Board Test), although models correctly identify the need for a rectangle to construct a complex figure, they fail to select sides of the appropriate length. These results indicate that while models possess some geometric recognition abilities, they struggle with accurately gauging line lengths and proportions, limiting their performance in tasks requiring precise spatial measurements.

While models frequently err in length and scale estimation across CF3 and VZ1 tasks, angular discrimination presents substantially more severe limitations. Systematic evaluation of directional vectors reveals a categorical bias toward canonical orientations: models consistently misclassified non-cardinal directions as 45-degree angles. In controlled testing with 20 systematically selected non-45-degree vectors (e.g., displacement vector [2, 1]), models achieve zero correct angular classifications, invariably defaulting to the nearest 45-degree approximation. This suggests that models possess coarse categorical representations of spatial orientation rather than continuous angular perception.

5. Related Work

Evaluation with Natural Images. Natural images are commonly used to evaluate the visual capabilities of MLLMs, as they more closely reflect real-world scenarios (Zhao et al., 2024; Liu et al., 2024a; Chow et al., 2025; Wadhawan et al., 2024). Recent research has emphasized MLLMs’ spatial reasoning abilities (Kamath et al., 2023; Liu et al., 2023), including tasks such as top-view map interpretation (Li et al., 2024) and region-level depth reasoning (Cheng et al., 2024). However, we argue that natural images often introduce additional noise and variability, making them less suitable for assessing core visual competencies. While benchmarks such as Blink (Fu et al., 2024), MMT-Bench and HallusionBench (Ying et al., 2024; Guan et al., 2024), and CoreCognition (Li et al., 2025b) incorporate synthetic images for tasks like IQ testing, visual hallucination detection, and physical reasoning, their overall focus remains primarily on natural image settings.

Evaluation with Synthetic Images. Synthetic images have been widely employed to evaluate the fundamental visual reasoning capabilities of MLLMs (Rahmanzadehgervi et al., 2024; Wu et al., 2024a; Chollet et al., 2025). Prior work has leveraged tasks such as Raven’s Progressive Matrices (Zhang et al., 2024; Song et al., 2024) and the Logic Test from the Chinese Civil Service Examination (Song et al., 2025), which include puzzles conceptually related to our I3 task. VisualSphinx (Feng et al., 2025) further extends this line of work by generating puzzles structurally similar to RPMs. Mental Rotation Tests have also been frequently

used (Ramakrishnan et al., 2025; Song et al., 2024), aligning with the design of our S1 and S2 tasks. In addition, synthetic images have supported evaluations of MLLMs on mathematical reasoning problems (Lu et al., 2024; Wang et al., 2025a), including polygons (Rudman et al., 2025) and graph-based challenges (Babaiee et al., 2025). Our proposed **VISFACTOR** advances this direction by providing a more comprehensive evaluation framework for core visual abilities, including 20 tests, systematically grounded in factor analysis from cognitive science. Furthermore, we implement automatic generation for 12 tests, enabling unlimited training data and ensuring the long-term scalability of the benchmark through high-difficulty content.

Enhancing MLLMs’ Visual Ability. A range of strategies have been proposed to strengthen spatial reasoning in MLLMs, including generating intermediate steps (Li et al., 2025a; Wu et al., 2024b), drawing auxiliary lines (Meng et al., 2023; Hu et al., 2024), incorporating coordinates or depth cues (Liu et al., 2025b; Cai et al., 2024), and augmenting training sets with reasoning data (Shao et al., 2024). Our approach enables automatic generation of high-quality, difficulty-controlled test cases, offering effectively unlimited training data to enhance MLLMs’ visual reasoning.

Using Psychological Tests on AI. Recent studies have evaluated AI models from psychological perspectives, including behavioral analysis (Coda-Forno et al., 2024), personality (Huang et al., 2024b;a), emotion (Huang et al., 2024b), and mental disorder (Coda-Forno et al., 2023). Research has found advanced human-like abilities in AI models, including Theory-of-Mind abilities (Liu et al., 2024c; Liang et al., 2023; Huang et al., 2025) and role-playing abilities (Ng et al., 2024; Wang et al., 2024b; 2025b). Inspired from cognitive science, our work provides a comprehensive framework for evaluating foundational visual abilities.

6. Conclusion

We presented **VISFACTOR**, the first factor-grounded benchmark that transposes twenty vision-centric subtests from the *Factor-Referenced Cognitive Test* battery into an automated image–text setting. A systematic evaluation of twenty state-of-the-art MLLMs uncovers a striking gap: despite their prowess on holistic leaderboards, the best model attains only 25.19/100 on **VISFACTOR**, often performing near chance on tasks that human novices solve with ease. Chain-of-Thought lifts scores only marginally, indicating that the deficit is architectural rather than prompt-level.

Beyond exposing a missing substrate for genuine visual reasoning, these findings carry practical ramifications. Hallucinated perception in safety-critical applications, brittle spatial reasoning in robotics, and misaligned multimodal

feedback loops all trace back to weak foundational vision. Bridging this gap will likely require *curriculum-style pre-training* that interleaves psychometric micro-tasks with natural images, *embodied or 3-D data* that grounds spatial relations, and *factor-aligned loss functions* that explicitly target low-level perceptual skills. By releasing **VISFACTOR** and its controllable-difficulty generator, we aim to catalyze these research directions and provide a rigorous yardstick for the next generation of visuocognitive AI.

References

- Adcock, C. J. and Martin, W. A. Flexibility and creativity. *The Journal of General Psychology*, 85(1):71–76, 1971.
- Anthropic. Claude 3.5 sonnet. *Anthropic Blog Jun 20 2024*, 2024. URL <https://www.anthropic.com/news/clause-3-5-sonnet>.
- Anthropic. Claude 3.7 sonnet and claude code. *Anthropic Blog Feb 24 2025*, 2025a. URL <https://www.anthropic.com/news/clause-3-7-sonnet>.
- Anthropic. Introducing claude 4. *Anthropic Blog Mar 22 2025*, 2025b. URL <https://www.anthropic.com/news/clause-4>.
- Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., and Merhof, D. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*, 2023.
- Babaiee, Z., Kiasari, P., Rus, D., and Grosu, R. Visual graph arena: Evaluating visual conceptualization of vision and multimodal large language models. In *Forty-second International Conference on Machine Learning*, 2025.
- Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Buckley, T., A. Diao, J., Rajpurkar, P., Rodman, A., and K. Manrai, A. Multimodal foundation models exploit text to make medical image predictions. *arXiv preprint arXiv:2311.05591*, 2023.
- ByteDance. Introduction to techniques used in seed1.6. *ByteDance Seed Blog Jun 25 2025*, 2025. URL https://seed/bytedance.com/en/seed1_6.
- Cai, W., Ponomarenko, I., Yuan, J., Li, X., Yang, W., Dong, H., and Zhao, B. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024.
- Carroll, J. B. *Psychometric Tests As Cognitive Tasks: A New” Structure of Intellect.” Technical Report No. 4*. ERIC, 1974.

- Cattell, R. B. *Abilities: Their structure, growth, and action.* Houghton Mifflin, 1971.
- Chen, S., Guo, X., Li, Y., Zhang, T., Lin, M., Kuang, D., Zhang, Y., Ming, L., Zhang, F., Wang, Y., et al. Ocean-ocr: Towards general ocr application via a vision-language model. *arXiv preprint arXiv:2501.15558*, 2025.
- Cheng, A.-C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., and Liu, S. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37, 2024.
- Chollet, F., Knoop, M., Kamradt, G., Landers, B., and Pinkard, H. Arc-agc-2: A new challenge for frontier ai reasoning systems. *arXiv preprint arXiv:2505.11831*, 2025.
- Chow, W., Mao, J., Li, B., Seita, D., Guizilini, V., and Wang, Y. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Coda-Forno, J., Witte, K., Jagadish, A. K., Binz, M., Akata, Z., and Schulz, E. Inducing anxiety in large language models can induce bias. *arXiv preprint arXiv:2304.11111*, 2023.
- Coda-Forno, J., Binz, M., Wang, J. X., and Schulz, E. Cog-bench: a large language model walks into a psychology lab. In *Forty-first International Conference on Machine Learning*, 2024.
- Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., Dong, X., Zang, Y., Zhang, P., Wang, J., et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.
- Dye, N. W. and Very, P. S. Growth changes in factorial structure by age and sex. *Genetic Psychology Monographs*, 1968.
- Ekstrom, R. B. *Cognitive Factors: Some Recent Literature.* ERIC, 1973.
- Ekstrom, R. B. and Harman, H. H. *Manual for kit of factor-referenced cognitive tests*, 1976. Educational testing service, 1976.
- Ekstrom, R. B. et al. *Problems of Replication of Seven Divergent Production Factors.* Technical Report No. 5. ERIC, 1974.
- Ekstrom, R. B. et al. *An Attempt to Confirm Five Recently Identified Cognitive Factors.* Technical Report No. 8. ERIC, 1975.
- Feng, Y., Xu, Z., Jiang, F., Li, Y., Ramasubramanian, B., Niu, L., Lin, B. Y., and Poovendran, R. Visualsphinx: Large-scale synthetic vision logic puzzles for rl. *arXiv preprint arXiv:2505.23977*, 2025.
- Frederiksen, J. R. Cognitive factors in the recognition of ambiguous auditory and visual stimuli. *Journal of Personality and Social Psychology*, 7(1p2):1, 1967.
- Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N. A., Ma, W.-C., and Krishna, R. Blink: Multi-modal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.
- Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., 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*, pp. 14375–14385, 2024.
- Guilford, J. P. *The nature of human intelligence.* McGraw-Hill, 1967.
- Guilford, J. P. and Hoepfner, R. Sixteen divergent-production abilities at the ninth-grade level. *Multivariate Behavioral Research*, 1(1):43–66, 1966.
- Guilford, J. P. and Hoepfner, R. The analysis of intelligence. (*No Title*), 1971.
- Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.
- Harris, M. L. and Harris, C. W. A factor analytic interpretation strategy. *Educational and Psychological Measurement*, 31(3):589–606, 1971.
- Hettema, J. Cognitive abilities as process variables. *Journal of personality and social psychology*, 10(4):461, 1968.
- Hu, Y., Shi, W., Fu, X., Roth, D., Ostendorf, M., Zettlemoyer, L., Smith, N. A., and Krishna, R. Visual sketchpad: Sketching as a visual chain of thought for multi-modal language models. *Advances in Neural Information Processing Systems*, 37, 2024.
- Huang, J.-t., Jiao, W., Lam, M. H., Li, E. J., Wang, W., and Lyu, M. R. On the reliability of psychological scales on large language models. In *Proceedings of The 2024 Conference on Empirical Methods in Natural Language Processing*, 2024a.
- Huang, J.-t., Lam, M. H., Li, E. J., Ren, S., Wang, W., Jiao, W., Tu, Z., and Lyu, M. R. Apathetic or empathetic?

- evaluating LLMs' emotional alignments with humans. *Advances in Neural Information Processing Systems*, 37, 2024b.
- Huang, J.-t., Li, E. J., Lam, M. H., Liang, T., Wang, W., Yuan, Y., Jiao, W., Wang, X., Tu, Z., and Lyu, M. R. Competing large language models in multi-agent gaming environments. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- Kamath, A., Hessel, J., and Chang, K.-W. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, 2023.
- Kavukcuoglu, K. Gemini 2.5: Our most intelligent ai model. *Google Blog Mar 25 2025*, 2025. URL <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/>.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022.
- Künnapas, T. Figural reversal rate and personal tempo. *Scandinavian journal of psychology*, 10(1):27–32, 1969.
- Li, C., Zhang, C., Zhou, H., Collier, N., Korhonen, A., and Vulić, I. Topviewrs: Vision-language models as top-view spatial reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1786–1807, 2024.
- Li, C., Wu, W., Zhang, H., Xia, Y., Mao, S., Dong, L., Vulić, I., and Wei, F. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025a.
- Li, Y., Gao, Q., Zhao, T., Wang, B., Sun, H., Lyu, H., Hawkins, R. D., Vasconcelos, N., Golan, T., Luo, D., and Deng, H. Core knowledge deficits in multi-modal language models. In *Forty-second International Conference on Machine Learning*, 2025b.
- Liang, T., He, Z., Huang, J.-t., Wang, W., Jiao, W., Wang, R., Yang, Y., Tu, Z., Shi, S., and Wang, X. Leveraging word guessing games to assess the intelligence of large language models. *arXiv preprint arXiv:2310.20499*, 2023.
- Liu, F., Emerson, G., and Collier, N. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.
- Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T., and Griffiths, T. L. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *Forty-second International Conference on Machine Learning*, 2025a.
- Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024a.
- Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.-C., Liu, C.-L., Jin, L., and Bai, X. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024b.
- Liu, Y., Chi, D., Wu, S., Zhang, Z., Hu, Y., Zhang, L., Zhang, Y., Wu, S., Cao, T., Huang, G., et al. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. *arXiv preprint arXiv:2501.10074*, 2025b.
- Liu, Z., Anand, A., Zhou, P., Huang, J.-t., and Zhao, J. Interintent: Investigating social intelligence of llms via intention understanding in an interactive game context. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024c.
- Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024.
- Meng, F., Yang, H., Wang, Y., and Zhang, M. Chain of images for intuitively reasoning. *arXiv preprint arXiv:2311.09241*, 2023.
- Messick, S. and French, J. W. Dimensions of cognitive closure. *Multivariate behavioral research*, 10(1):3–16, 1975.
- Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta Blog Sep 25 2024*, 2024. URL <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>.

- MoonshotAI. Multimodal image understanding model moonshot-v1-vision-preview. *Moonshot AI Blogs Jan 2025*, 2025. URL <https://platform.moonshot.cn/docs/guide/use-kimi-vision-model>.
- Ng, M. T., Tse, H. T., Huang, J.-t., Li, J., Wang, W., and Lyu, M. R. How well can llms echo us? evaluating ai chatbots' role-play ability with echo. *arXiv preprint arXiv:2404.13957*, 2024.
- OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. *OpenAI Blog Jul 18 2024*, 2024. URL <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
- OpenAI. Introducing gpt-4.1 in the api. *OpenAI Blog Apr 14 2025*, 2025a. URL <https://openai.com/index/gpt-4-1/>.
- OpenAI. Introducing openai o3 and o4-mini. *OpenAI Blog Apr 16 2025*, 2025b. URL <https://openai.com/index/introducing-o3-and-o4-mini/>.
- Pawlick, K. Concepts and calculations in human cognitive abilities. *Cattell, RB (Ed.), Handbook of multivariate experimental psychology*, 1966.
- Peng, S., Fu, D., Gao, L., Zhong, X., Fu, H., and Tang, Z. Multimath: Bridging visual and mathematical reasoning for large language models. *arXiv preprint arXiv:2409.00147*, 2024.
- Petrov, Y. Memory structure as a psychic function. *Voprosi Psichologii*, 16:132–136, 1970.
- Rahmanzadehgervi, P., Bolton, L., Taesiri, M. R., and Nguyen, A. T. Vision language models are blind: Failing to translate detailed visual features into words. *arXiv preprint arXiv:2407.06581*, 2024.
- Ramakrishnan, S. K., Wijmans, E., Kraehenbuehl, P., and Koltun, V. Does spatial cognition emerge in frontier models? In *The Thirteenth International Conference on Learning Representations*, 2025.
- Roff, M. A factorial study of tests in the perceptual area. *Psychometric Monograph*, (8), 1953.
- Royce, J. *The conceptual framework for a multifactor theory of individuality. In " Multivariate Analysis and Psychological Theory"(JR Royce, ed.)*. Academic Press, London and New York, 1973.
- Rudman, W., Golovanevsky, M., Bar, A., Palit, V., LeCun, Y., Eickhoff, C., and Singh, R. Forgotten polygons: Multimodal large language models are shape-blind. *arXiv preprint arXiv:2502.15969*, 2025.
- Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., and Li, H. Visual cot: Advancing multimodal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.
- Shepard, R. N. and Feng, C. A chronometric study of mental paper folding. *Cognitive psychology*, 3(2):228–243, 1972.
- Shepard, R. N. and Metzler, J. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971.
- Song, W., Li, Y., Xu, J., Wu, G., Ming, L., Yi, K., Luo, W., Li, H., Du, Y., Guo, F., et al. M3gia: A cognition inspired multilingual and multimodal general intelligence ability benchmark. *arXiv preprint arXiv:2406.05343*, 2024.
- Song, Y., Ou, T., Kong, Y., Li, Z., Neubig, G., and Yue, X. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025.
- Sprague, Z., Yin, F., Rodriguez, J. D., Jiang, D., Wadhwa, M., Singhal, P., Zhao, X., Ye, X., Mahowald, K., and Durrett, G. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Team, Q. Introducing qwen-vl. *Qwen Blogs Jan 2024*, 2024. URL <https://qwenlm.github.io/blog/qwen-vl/>.
- Thurstone, L. L. Primary mental abilities:. *Psychology Monographs*, (1), 1938.
- Thurstone, L. L. *A factorial study of perception*. The University of Chicago Press, 1944.
- Thurstone, L. L. Theories of intelligence. *The scientific monthly*, 62(2):101–112, 1946.
- Wadhawan, R., Bansal, H., Chang, K.-W., and Peng, N. Contextual: Evaluating context-sensitive text-rich visual reasoning in large multimodal models. In *Forty-first International Conference on Machine Learning*, 2024.
- Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.
- Wang, P., Li, Z.-Z., Yin, F., Ran, D., and Liu, C.-L. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19541–19551, 2025a.

- Wang, X., Xiao, Y., Huang, J.-t., Yuan, S., Xu, R., Guo, H., Tu, Q., Fei, Y., Leng, Z., Wang, W., et al. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1840–1873, 2024b.
- Wang, X., Wang, H., Zhang, Y., Yuan, X., Xu, R., Huang, J.-t., Yuan, S., Guo, H., Chen, J., Wang, W., et al. Coser: Coordinating llm-based persona simulation of established roles. *arXiv preprint arXiv:2502.09082*, 2025b.
- Wardell, D. Possible changes in the taxonomies in royce. *Center for Advanced Study in Theoretical Psychology*, pp. 252–261, 1973.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022.
- Werdelin, I. and Stjernberg, G. The relationship between difficulty and factor loadings of some visual-perceptual tests. *Scandinavian Journal of Psychology*, 12(1):21–28, 1971.
- Witkin, H. A. *A manual for the embedded figures tests*. Consulting Psychologists Press, 1971.
- Wu, A., Brantley, K., and Artzi, Y. A surprising failure? multimodal llms and the nlvr challenge. *arXiv preprint arXiv:2402.17793*, 2024a.
- Wu, W., Mao, S., Zhang, Y., Xia, Y., Dong, L., Cui, L., and Wei, F. Mind’s eye of llms: visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems*, 37: 90277–90317, 2024b.
- Yang, Z., Chen, J., Du, Z., Yu, W., Wang, W., Hong, W., Jiang, Z., Xu, B., Dong, Y., and Tang, J. Mathglm-vision: Solving mathematical problems with multi-modal large language model. *arXiv preprint arXiv:2409.13729*, 2024.
- Ying, K., Meng, F., Wang, J., Li, Z., Lin, H., Yang, Y., Zhang, H., Zhang, W., Lin, Y., Liu, S., et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. In *Forty-first International Conference on Machine Learning*, 2024.
- Zhang, Y., Bai, H., Zhang, R., Gu, J., Zhai, S., Susskind, J., and Jaitly, N. How far are we from intelligent visual deductive reasoning? In *The First Conference on Language Modeling*, 2024.
- Zhao, H. H., Zhou, P., Gao, D., and Shou, M. Z. Lova3: Learning to visual question answering, asking and assessment. *Advances in Neural Information Processing Systems*, 37, 2024.
- Zimmerman, W. S. The influence of item complexity upon the factor composition of a spatial visualization test. *Educational and Psychological Measurement*, 14(1):106–119, 1954.

A. Descriptions and Prompts for All Subtests

This section introduces each subtest in detail and provides the prompts we use in **VISFACTOR**.

A.1. Closure Flexibility (CF)

The Factor:

"The ability to hold a given visual percept or configuration in mind so as to disembed it from other well defined perceptual material."

Flexibility of closure, a cognitive factor involving the identification of a configuration within a distracting perceptual field, has been linked to the concept of field independence, though they are not considered identical constructs. [Witkin \(1971\)](#) related this factor to both Thurstone's flexibility of closure ([Thurstone, 1938](#)) and Guilford's adaptive flexibility ([Guilford, 1967](#)), suggesting similarities to field independence. [Royce \(1973\)](#) proposed that flexibility of closure may interact with higher-order cognitive factors, while [Hettema \(1968\)](#) posited it as conceptually situated between flexibility and speed of closure. [Wardell \(1973\)](#) argued for its identity with figural adaptive flexibility. [Carroll \(1974\)](#) defined flexibility of closure as involving short-term memory processes that match a figure to its surrounding field, and [Cattell \(1971\)](#) framed it as a restructuring ability central to personality and practical intelligence.

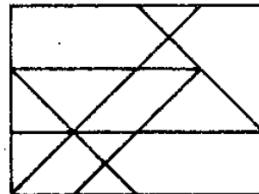
Prompt for CF1: Hidden Figures Test

Look at the two images:

Below is the first image, one simple shape:



Below is the second image, a larger, complex pattern:



Task: Decide whether the shape in the first image is hidden anywhere inside the second image. The shape will never be rotated, flipped, or resized. The shape will always be right-side-up and exactly the same size as in the first image.

Output: Respond with only one word: "TRUE" if it is present, "FALSE" if it is not, in JSON format as follows: {"answer": YOUR_ANSWER_HERE}.

Prompt for CF2: Hidden Patterns Test

Look at the two images:

Below is the first image, a model:



Below is the second image, a pattern:



Task: Decide if the model in the first image is hidden anywhere in the pattern in the second image. The model must be in that exact position, no turning or flipping.

Output: Respond with only one word: "TRUE" if it is present, "FALSE" if it is not, in JSON format as follows: {"answer": YOUR_ANSWER_HERE}.

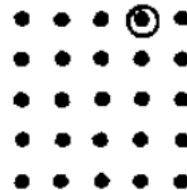
Prompt for CF3: Copying Test

Look at the two images:

Below is the first image, a simple line shape:



Below is the second image, a 5 times 5 grid of dots; one dot is circled as the starting point:



Task: Begin at the circled dot on the second image. Copy the shape shown in the first image onto the grid so that every corner of the line sits exactly on a dot. When you are done, the pattern on the grid must look the same as the shape in the first image.

Output: Respond with only a tuple, the dot you finally reach, as a (row, column) pair where the row is counted top-to-bottom and the column left-to-right, in JSON format as follows: {"answer": YOUR_ANSWER_HERE}.

A.2. Closure Speed (CS)

The Factor:

“The ability to unite an apparently disparate perceptual field into a single concept.”

The concept of speed of closure refers to the ability to rapidly recognize and organize ambiguous or partially obscured visual stimuli, a process distinct from flexibility of closure, which involves identifying a known configuration within complex figures. This skill is associated with the early identification of out-of-focus and close-up images (Frederiksen, 1967), and involves long-term memory search strategies (Carroll, 1974). It has been linked to cognitive factors like restraint-timidity (Cattell, 1971) and may reflect a broader aptitude for visual scanning and cognitive-affective integration (Thurstone, 1944; Wardell, 1973; Roff, 1953; Adcock & Martin, 1971; Messick & French, 1975).

Prompt for CS3: Snowy Pictures

Look at this image below:



Task: Even if parts are hidden, name the main object you see.

Output: Respond with only one or two words, in JSON format as follows: {"answer": "YOUR_ANSWER_HERE"}.

Prompt for CS1: Gestalt Completion Test

Look at the incomplete drawing below:

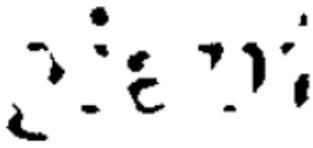


Task: Write the name of the object you think it shows.

Output: Respond with only one or two words, in JSON format as follows: {"answer": "YOUR_ANSWER_HERE"}.

Prompt for CS2: Concealed Words Test

Look at the image below, which shows one lowercase English word, but parts of the letters are missing:



Task: Write the complete word. The word is at least four letters long. Use only lowercase letters.

Output: Respond with only the answer word, in JSON format as follows: {"answer": "YOUR_ANSWER_HERE"}.

A.3. Induction (I)

The Factor:

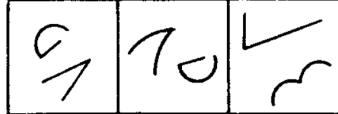
“The reasoning abilities involved in forming and trying out hypotheses that will fit a set of data.”

Research on inductive reasoning suggests it involves both concept formation and hypothesis testing, functioning as a synthesizing process (Wardell, 1973). Evidence points to several subfactors, with figure classification being particularly distinct (Harris & Harris, 1971). Guilford & Hoepfner (1966) identified 16 types of inductive ability, while Dye & Very (1968) proposed distinct inductive and symbolic-inductive reasoning factors. Though Pawlick (1966) argued that induction and general reasoning are not separate, Catell (1971) allowed for a possible figural reasoning factor. Carroll (1974) emphasized the role of long-term memory search in induction, noting that success depends on the content of a “general logic store” and the ability to construct new hypotheses through serial operations.

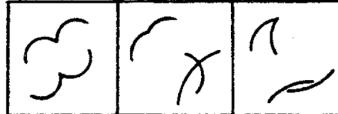
Prompt for I3: Figure Classification

Look at the four images:

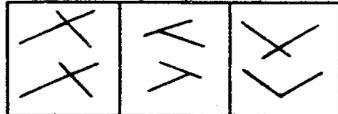
Below is the first image, three figures in the Group 1:



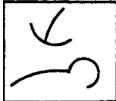
Below is the second image, three figures in the Group 2:



Below is the third image, three figures in the Group 3:



Below is the fourth image, the figure to classify:



Task: Inside a group, all three figures share one rule. Different groups follow different rules. Find the rule and decide whether the figure in the fourth image belongs to Group 1, 2, or 3.

Output: Respond with only the group number (1, 2, or 3), in JSON format as follows: {"answer": YOUR_ANSWER_HERE}.

A.4. Associative Memory (MA)

The Factor:

“The ability to recall one part of a previously learned but otherwise unrelated pair of items when the other part of the pair is presented.”

Tasks assessing this factor are similar to those used in paired-associates learning and may involve memory for non-meaningful material. This factor reflects intermediate-term memory processes, where individual differences arise from the use of strategies such as short-term rehearsal and the identification of mnemonic mediators in long-term memory (Carroll, 1974).

Prompt for MA1: Picture-Number Test

Look at the two images:

Below is the first image, the 21 picture-number pairs to memorize:

	34		46		55
	73		69		10

Below is the second image, a picture:



Task: Write down the number that the picture in the second image belongs to, as shown in the first image.

Output: Respond with only a number, in JSON format as follows:
{"answer": YOUR_ANSWER_HERE}.

A.5. Visual Memory (MV)

The Factor:

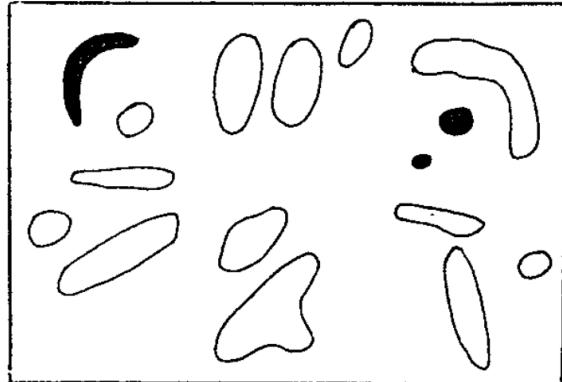
“The ability to remember the configuration, location, and orientation of figural material.”

Visual memory involves distinct cognitive processes beyond mere test content, as suggested by research on iconic memory, which stores visual impressions (Thurstone, 1946). While Thurstone (1946) argued that “the memorizing factor transcends the nature of the content,” later studies demonstrated that visual memory is a multifaceted construct. Guilford (1967) identified six figural memory abilities, and Petrov (1970) distinguished between factors for iconic memory and short-term visual retention, indicating the presence of sub-factors within visual memory.

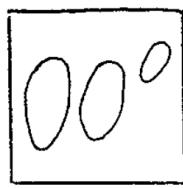
Prompt for MV1: Shape Memory Test

Look at the two images:

Below is the first image, memorize each shape and the way it is turned:



Below is the second image:



Task: Decide whether the following statement is true or false: the second image does not show any part of the first image with the same shapes in the same orientation.

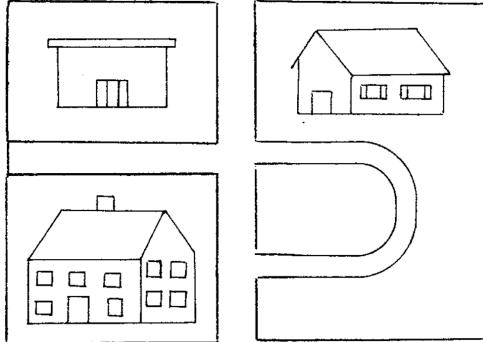
(!!!) Three other prompts are: (1) the second image does not show any part of the first image with the same shapes in the same orientation (2) some part of the first image contains the second image with the same shapes in the same orientation (3) some part of the first image does not contain the second image with the same shapes in the same orientation

Output: Respond with only one word: “TRUE” or “FALSE”, in JSON format as follows: {"answer": "YOUR_ANSWER_HERE"}.

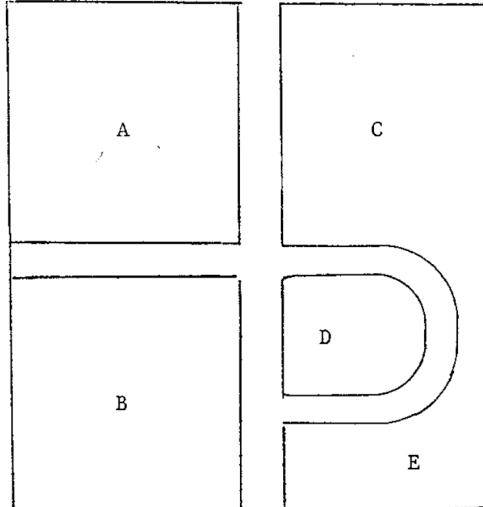
Prompt for MV2: Building Memory

Look at the two images:

Below is the first image, memorize where every building sits on this street map:



Below is the second image, the streets are the same, but each block is labeled A, B, C, D, E:



Below is the third image, a building:



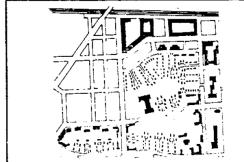
Task: Decide whether the building in the third image is in block E.

Output: Respond with only one word: “TRUE” if it is, “FALSE” if it is not, in JSON format as follows: {"answer": "YOUR_ANSWER_HERE"}.

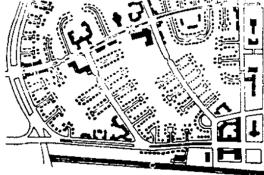
Prompt for MV3: Map Memory

Look at the two images:

Below is the first image, memorize each map:



Below is the second image, a single map:



Task: Decide whether the following statement is true or false: the map in the second image appears in the first image.

(!!!) Three other prompts are: (1) the map in the second image does not appear in the first image (2) the maps in the first image contain the map in the second image (3) the maps in the first image do not contain the map in the second image

Output: Respond with only one word: "TRUE" or "FALSE", in JSON format as follows: {"answer": YOUR_ANSWER_HERE}.

A.6. Perceptual Speed (P)

The Factor:

"Speed in comparing figures or symbols, scanning to find figures or symbols, or carrying out other very simple tasks involving visual perception."

Perceptual speed has been described as comprising three components: (1) perceptual fluency, or the readiness with which individuals switch between alternating percepts; (2) decision speed, or the readiness of choice when the response is not fully driven by sensory input (Thurstone, 1938; Künnapas, 1969); and (3) immediate perceptual memory. Carroll (1974) defines perceptual speed as involving the temporal aspects of visual search through a field of specified elements by accessing sensory buffers. It may be related to flexibility of closure (Pawlick, 1966; Ekstrom, 1973) or to an "automatic process" factor. Additionally, (Royce, 1973) suggested it may be a subfactor of the scanning cognitive style and possibly linked to the automatization cognitive style. It may be the centroid of several subfactors (including form discrimination and symbol discrimination) which can be separated but are more usefully treated as a single concept for research purposes.

Prompt for P3: Identical Pictures Test

Look at the two images:

Below is the first image, the target object:



Below is the second image, the test object:



Task: Decide whether the two objects are exactly the same.

Output: Respond with only one word: "TRUE" if they are, "FALSE" if they are not, in JSON format as follows: {"answer": YOUR_ANSWER_HERE}.

A.7. Logical Reasoning (RL)

The Factor:

“The ability to reason from premise to conclusion, or to evaluate the correctness of a conclusion.”

The cognitive factor historically referred to as “Deduction” (Thurstone, 1938), later termed “Syllogistic Reasoning,” and also known as “Logical Evaluation”, involves evaluating the correctness of presented answers rather than pure deductive reasoning (Guilford, 1967). Carroll (1974) emphasized its complexity, highlighting the need for retrieving meanings and algorithms from long-term memory and applying serial operations, with individual differences influenced by content, timing, and attentional focus on stimuli.

Prompt for RL2: Diagramming Relationships

Look at the image below:



Each circle stands for one group of things. Simple rules:

1. A circle inside another: all things in the inner group belong to the outer group.
2. Circles that overlap partly: the two groups share some, but not all, things.
3. Circles that do not touch: the two groups share nothing.

Task: Decide whether the image follows these rules for the three groups: Desks, furniture, pencils.

Output: Respond with only one word: “TRUE” if it shows the relationships for the three groups, “FALSE” if it does not, in JSON format as follows: { "answer": YOUR_ANSWER_HERE }.

Prompt for S1: Card Rotations Test

Look at the two images:

Below is the first image, the target shape:



Below is the second image, the test shape:



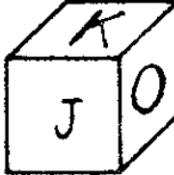
Task: The test shapes may be rotated, but they are not allowed to be flipped (mirrored). Decide whether test shape is the same shape as the target.

Output: Respond with only one word: “TRUE” if it is, “FALSE” if it is not, in JSON format as follows: { "answer": YOUR_ANSWER_HERE }.

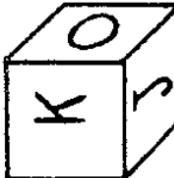
Prompt for S2: Cube Comparisons Test

Look at the two images:

Below is the first image, the first cube:



Below is the second image, the second cube:



Rules:

1. Each cube has six faces. Every face shows a different letter, number, or symbol.
2. Hidden faces may show any symbols, but no symbol appears on more than one face of the same cube.

Task: Decide whether the following statement is true or false: the first cube is a certain view of the second cube after it is turned.

(!!!) Three other prompts are: (1) the first cube is not any view of the second cube no matter how it is turned (2) the second cube is a certain view of the first cube after it is turned (3) the second cube is not any view of the first cube no matter how it is turned

Output: Respond with only one word: “TRUE” or “FALSE”, in JSON format as follows: { "answer": YOUR_ANSWER_HERE }.

A.8. Spatial Relations (S)

The Factor:

“The ability to perceive spatial patterns or to maintain orientation with respect to objects in space.”

Research has differentiated between spatial orientation and visualization, suggesting that while spatial orientation involves perceiving figures as wholes and performing mental rotation (Zimmerman, 1954; Werdelin & Stjernberg, 1971), visualization requires more complex restructuring and serial operations (Carroll, 1974; Shepard & Metzler, 1971). Although some distinguished between spatial relations and orientation (with the latter involving the observer’s body), Guilford & Hoepfner (1971) treated them as a single cognitive factor linked to egocentrism.

A.9. Spatial Scanning (SS)

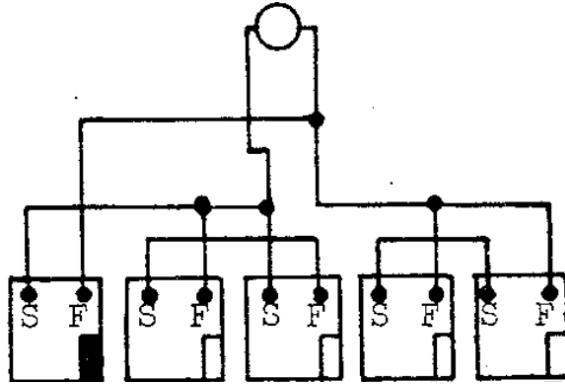
The Factor:

“Speed in exploring visually a wide or complicated spatial field.”

The ability to navigate a paper maze relies on quickly scanning for viable paths and rejecting false leads, engaging a visual search process somewhat akin to scanning text for comprehension. While sometimes associated with “planning,” the process primarily reflects a willingness to visually evaluate options before committing. Carroll (1974) noted that this skill involves managing sensory input and that individuals may adopt strategies such as working backward from the goal to simplify the task.

Prompt for SS2: Choosing A Path

Look at the diagram shown in the image below:



Rules:

1. You may switch lines only where a black dot is drawn.
2. Lines that cross or touch without a dot are not connected.
3. The path must stay inside the chosen box and must not stop at a dead-end.

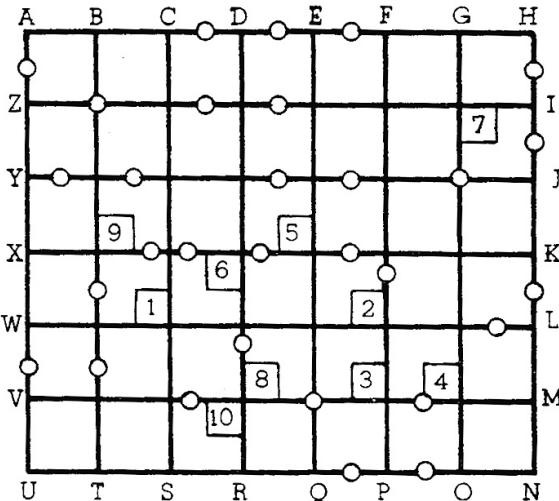
Task: For box E, decide if there is one continuous line that:

1. Starts at S inside that box.
2. Reaches the single circle at the top.
3. Comes back to F inside the same box without entering any other box.

Output: Respond with only one word: “TRUE” if box E meets all the rules. “FALSE” if it does not, in JSON format as follows: {"answer": YOUR_ANSWER_HERE}.

Prompt for SS3: Map Planning Test

Look at the city map shown in the image below:



In the map:

1. Streets = black lines.
2. Circles = road-blocks (you cannot cross there).
3. Numbered squares = buildings.

Task: Find the shortest street route from F to T. Rules:

1. The route will always touch the side of one and only one numbered building.
2. Touching only a corner does not count.
3. Move only along streets (horizontal or vertical), never through circles.

Output: Respond with only one number: the number on the building your shortest route touches, in JSON format as follows: {"answer": YOUR_ANSWER_HERE}.

A.10. Visualization (VZ)

The Factor:

“The ability to manipulate or transform the image of spatial patterns into other arrangements.”

Visualization and spatial orientation are related cognitive factors, yet visualization involves mentally restructuring figures into components for manipulation, making it more complex than spatial orientation, which deals with rotating entire figures. While some researchers view visualization as a higher-order or secondary factor encompassing various spatial abilities (Cattell, 1971; Royce, 1973), others emphasize its reliance on short-term visual memory and serial processing (Carroll, 1974). Analytic strategies, such as identifying symmetry and reflection planes, are often used in visualization tasks, as illustrated by Shepard & Feng (1972)’s work on paper-folding tests.

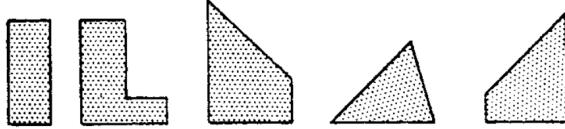
Prompt for VZ1: Form Board Test

Look at the two images:

Below is the first image, which is the figure you must make:



Below is the second image, which are the five pieces you can use:



Rules:

1. Use 2–5 of the pieces to fill the figure exactly.
2. You may rotate pieces but do not flip them.

Task: Decide whether the Fifth piece is in the set of pieces that makes the figure.

Output: Respond with only one word: “TRUE” if it is or “FALSE” if it is not, in JSON format as follows: {"answer": YOUR_ANSWER_HERE}.

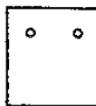
Prompt for VZ2: Paper Folding Test

Look at the two images:

Below is the first image, a step-by-step drawing of a square sheet being folded (solid lines) and then punched (small circle marks):



Below is the second image, the same sheet shown completely unfolded, with any holes that appear:



Task:

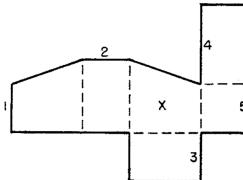
1. Mentally follow every fold in the first image exactly as drawn. Do not flip or rotate the paper except for the folds shown.
2. Imagine a hole being punched through all layers where each circle is drawn.
3. Unfold the paper, step by step, in reverse order of the folds, keeping the sheet’s original orientation.
4. After it is flat, note where every hole should appear on the sheet.
5. Compare this mental result with the pattern of holes in the second image.

Output: Respond with only one word: “TRUE” if every hole (number and position) in the second image matches your mental result exactly, otherwise “FALSE”, in JSON format as follows: {"answer": YOUR_ANSWER_HERE}.

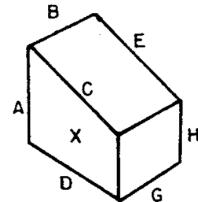
Prompt for VZ3: Surface Development Test

Look at the two images:

Below is the first image, the flat paper:



Below is the second image, the 3-D object:



Task: Fold the flat paper in the first image on every dashed line so that the face marked X ends up on the outside of the 3-D object in the second image. Decide edge 5 on the flat paper in the first image touches which lettered edge on the 3-D object in the second image after folding.

(!!) Decide whether the pair of one letter on the 3-D object in the second image and one number on the flat paper in the first image: (5, H) are two edges that touch each other after folding.

Output: Respond with only one letter, in JSON format as follows: {"answer": YOUR_ANSWER_HERE}.

(!!) Output: Respond with only one word: “TRUE” if they touch, “FALSE” if they do not, in JSON format as follows: {"answer": YOUR_ANSWER_HERE}.

B. Limitations

Psychometric Purity. **ViSFACTOR** inherits the FRCT assumption that each sub-test isolates a single latent visual factor (Ekstrom et al., 1974; 1975). In reality, human cognition is highly inter-dependent: even a seemingly “pure” mental-rotation item also taps working memory, executive control, and verbal encoding. Sub-test scores should therefore be read as *upper-bound indicators* of factor competence, not as proofs of modularity. For the same reason, factor-level comparisons across models must be interpreted with caution, especially when subtle prompt differences can shift the mixture of underlying skills that a model exploits.

Digitization Gap. The original FRCT was administered on paper, under timed and proctored conditions. Our pipeline converts items to separate images and accepts typed responses, eliminating motor demands but also removing contextual cues such as page layout and time pressure. We also simplify the original instructions for MLLMs. These changes inevitably alter item difficulty, so direct numerical comparisons with legacy human norms are inappropriate.

Missing Contemporary Human Baseline. We have not re-collected human performance under the digital protocol, leaving open questions about the relative difficulty of the adapted items and about ceiling effects that may mask model progress. Gathering calibrated human baselines—ideally across age groups and devices—would help normalize model scores and identify items whose difficulty distribution shifted during digitization.