FLUSHPuzzle: Fine-grained Logical Understanding through Structural Visual Reasoning Benchmark

Anonymous ACL submission

Abstract

Despite the rapid advancement of Multimodal Large Language Models (MLLMs), their reasoning capabilities are often constrained by perceptual fragility and a lack of transparent logical derivation. This frequently leads to cascaded failures, where minor perceptual inaccuracies propagate through the reasoning chain. We propose a novel, automated rule-based generation framework **FLUSH-Gen** that ensures rigorous logical methodology consistency by decoupling visual synthesis from visual attributes. Leveraging this framework, we introduce **FLUSHPuzzle**, a hierarchical 20,000 instances benchmark comprising 30 perception primitives and 200 reasoning subclasses. Unlike existing benchmarks, each of our 20,000 samples is paired with a verifiable reasoning trace explicitly mapped to low-level visual elements, enabling fine-grained diagnostic evaluation. Our experiments demonstrate that fine-tuning 8B-parameter models in the FLUSH-Puzzle train set yields significant performance gains, achieving an absolute accuracy improvement of 15.8% and competitiveness with proprietary models such as Gemini 3 Pro.

"Poker is a game of incomplete information. You have to use logic to fill the gaps the eyes cannot see."

1 Introduction

Although Multimodal Large Language Models (MLLMs) perform strongly on general tasks such as dialog and captioning (Liu et al., 2023), they often struggle with structured visual reasoning that requires consistent spatio-temporal representations and well-defined transformation rules (Tang et al., 2025b; Yue et al., 2024). We argue that these failures come largely from unstable perception. Errors in model prediction during early visual understanding propagate through the reasoning process, making it difficult to determine whether the shortcomings of a model are perceptual or logical (Lu et al., 2023). In abstract puzzle settings, instance-level accuracy provides limited diagnostic insight (Li et al., 2025), which motivates our hierarchical framework that explicitly links Perception Primitives to problem design to enable attributable analysis.

Data reliability remains a major challenge because many existing automated benchmarks like VisualSphinx (Feng et al., 2025) use MLLMs to generate both images and their accompanying instructions. This tight coupling can introduce logical hallucinations in which visual elements no longer align with the underlying reasoning rules, thus injecting uncontrolled noise and compromising the validity of the training (Tang et al., 2025a). To mitigate this issue, we adopt a verifiable rule-based pipeline that strictly separates visual synthesis from visual attributes, while using MLLMs only for open-ended linguistic polishing. This design preserves expressive diversity without sacrificing logical rigor (Wei et al., 2025; Huang et al., 2025).

Moreover, most current benchmarks provide only final answers and lack a fine-grained correspondence between intermediate reasoning steps and low-level perceptual units. Without such grounded explanations, it becomes challenging to disentangle whether model errors stem from defective perceptual processing or flawed rule application, a critical gap that impedes the design of targeted intervention strategies (Lu et al., 2022). FLUSHPuzzle addresses this limitation by providing reasoning traces that are explicitly aligned with perception primitives, enabling precise diagnosis and providing robust supervision signals (Wei et al., 2022). Based on FLUSH-Gen, an automated system that leverages 30 primitives and 200 subclasses, we investigate how data quality and grounded training influence MLLM performance through rigorous analysis.

Our main contributions are as follows:

- **FLUSH-Gen:** A rule-based framework that ensures logical integrity by decoupling visual

synthesis from cognitive logic, supporting attributable analysis.

- **FLUSHPuzzle Dataset:** A hierarchical benchmark featuring 30 foundational perception primitives and 200 reasoning subclasses, paired with verifiable reasoning traces.
- **Empirical Insights:** Evidence that training on our grounded data substantially improves MLLM reasoning while revealing findings such as the strategy collapse phenomenon.

2 Related Work

This section reviews the perceptual capabilities of multimodal large language models and surveys representative benchmarks for visual puzzle reasoning.

2.1 Perceptual Capabilities of MLLMs

Early multimodal models achieved strong performance on tasks such as image captioning and visual question answering by integrating visual encoders with language models (Liu et al., 2023). However, these tasks often primarily evaluate basic object recognition and do not rigorously probe deeper visual reasoning. Recent studies show that tasks requiring fine-grained perception and spatial understanding remain challenging for current MLLMs (Tang et al., 2025b; Yue et al., 2024). For example, the BLINK benchmark reports a substantial performance gap, with even advanced models performing near chance on perception-intensive tasks such as depth estimation and viewpoint consistency (Fu et al., 2024). One reason behind this phenomenon is that many MLLMs initially transform images into simplified textual descriptions or representations, a process that can eliminate essential geometric information (such as symmetry and rotation). While some approaches add vision-focused objectives to strengthen low-level perception, they often fail to clearly distinguish between errors in perception and mistakes in subsequent reasoning. We resolve this ambiguity by explicitly tying each reasoning step to the specific visual evidence it relies on, creating a transparent basis that promotes stable perception before any logical rules are applied.

2.2 MLLM Puzzle Reasoning Benchmarks

Existing benchmarks such as VisuLogic (Xu et al., 2025) and VisualPuzzles (Song et al., 2025) evaluate vision-centric reasoning using human-verified problems or materials adapted from civil service examinations. However, these datasets generally assess only the final responses and omit intermediate reasoning traces. This end-point evaluation approach impedes attributable analysis, leaving it uncertain whether errors stem from perceptual issues (e.g., misinterpreting a visual element) or from failures in multi-step logical reasoning. In addition, dependency on fixed exam sets can restrict systematic variation in the visual stimuli when compared with the more fine-grained design space enabled by our framework.

Recent efforts such as VisualSphinx V1 (Feng et al., 2025) and LogicVista (Xiao et al., 2024) incorporate explanations, but they face challenges in logical reliability and task purity. VisualSphinx relies on MLLMs for automated generation, which can introduce logical hallucinations when generated rules or descriptions deviate from the true image structure and ground truth, thereby injecting uncontrolled noise that undermines both training and evaluation. LogicVista provides human-written explanations, but its inclusion of text-heavy or OCR-dependent questions makes it difficult to isolate purely visual reasoning. Similarly, although BLINK (Fu et al., 2024) includes pattern-completion tasks that reduce linguistic mediation, it does not expose the underlying reasoning process. FLUSH-Gen addresses these limitations by introducing an automated generation system and a hierarchical design space that combines 30 foundational Perception Primitives with 200 puzzle subclasses. By avoiding MLLM-based generation and enforcing deterministic alignment between reasoning steps and distinct perceptual units via rule-based construction, our approach promotes logical consistency and stabilizes the perceptual layer. This design enables precise diagnosis of whether a model misidentifies a geometric shape or misapplies a transformation rule, yielding a more reliable benchmark for complex graphic reasoning.

3 FLUSH-Gen

As shown in Figure 1, our FLUSH-Gen for generating visual reasoning problems adopts a modular architecture that produces datasets with strong logical consistency and substantial visual diversity. By separating the pipeline into (i) the Perception Primitive Synthesis module and (ii) the Hierarchical Problem Synthesis module, the framework enables end-to-end control over both pixel-level rendering and higher-level reasoning structure. This design grounds each generated sample in verifiable rules while preserving the complexity required for rigorous multimodal evaluation.

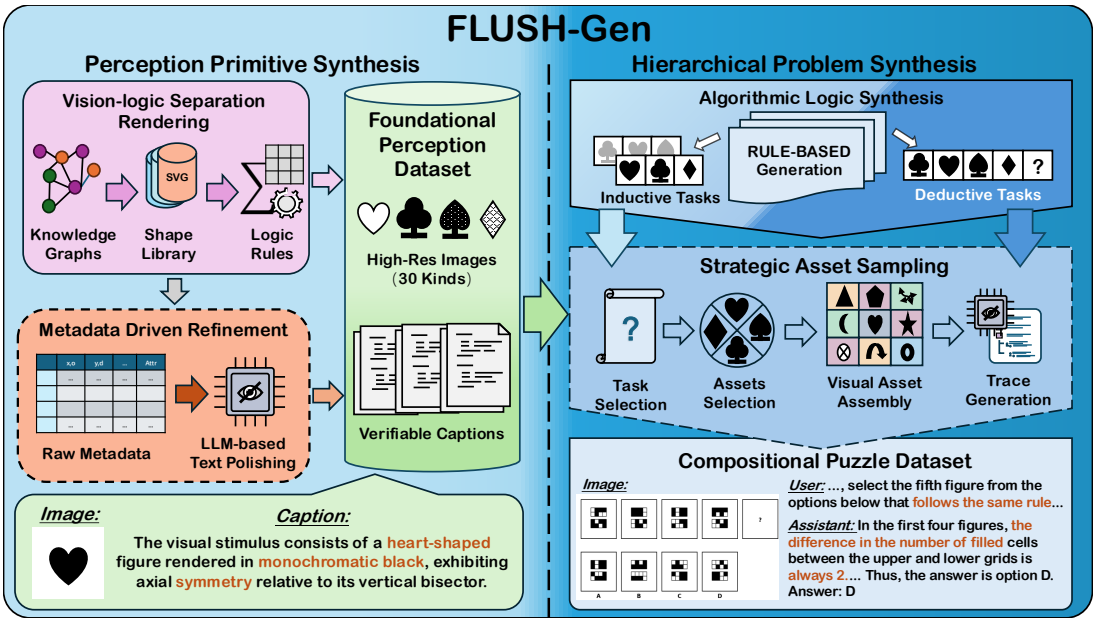


Figure 1: Overview of the FLUSH-Gen framework for rule-based data synthesis. The system ensures logical integrity by decoupling (i) the perception primitive synthesis from (ii) the hierarchical problem synthesis.

3.1 Perception Primitive Synthesis

The Perception Primitive Synthesis module in the FLUSH-Gen framework forms the foundation for synthesizing high-fidelity visual assets and their corresponding semantic descriptions. At this stage, we use a deterministic grounding procedure to align visual signals with their linguistic representations.

Vision-logic Separation Rendering. The perception engine follows a Vision-logic Separation Rendering principle to improve precision. As illustrated in the pink area of Figure 1, the framework distills visual attributes e.g., number and orientation of axis of symmetry) into structured Knowledge Graphs, where the visual attributes form nodes in the graph. These nodes can be instantiated into visual assets via Shape Library, such as SVG. Applying deterministic Logic Rules orchestrates these assets into complex configurations, producing the Foundational Perception Dataset with High Resolution Images across 30 categories. This separation maintains consistent quality while reducing rendering noise for multimodal inference.

Metadata Driven Refinement. To ensure multimodal supervision, the Metadata Driven Refinement module constructs descriptions from deterministic facts. As shown in the orange area of Figure 1, Raw Metadata captures instance-level ground truth including coordinates (x, y), visual attributes, etc. To enhance linguistic variety, Blind LLM Polishing takes the raw metadata, which are structured tabular data, as input and refines it into a human-friendly caption. The refinement process is achieved in a text-based LLM without image access to prevent hallucination from the visual side, potentially. This process yields finalized Verifiable Captions as the ground truth for the foundational perception data.

3.2 Hierarchical Problem Synthesis

Within the FLUSH-Gen framework, the Hierarchical Problem Synthesis module functions as the logical engine that orchestrates the algorithmic composition of puzzles. By strategically mapping symbolic rule sets to visual layouts, the module guarantees that the synthesized problem adheres to rigorous cognitive paradigms.

Algorithmic Logic Synthesis. As illustrated in the top right of Figure 1, the process starts with Algorithmic Logic Synthesis to define the reasoning structure. For Inductive Tasks, the system pairs image sets under mutually exclusive geometric rules, while for Deductive Tasks, it derives premises from spatial patterns and introduces adversarial distractors. This stage utilizes RULE-BASED axioms to simultaneously generate raw Reasoning Traces. This symbolic foundation provides the foundation that logical steps are verifiable and decoupled from specific visual instances before final assembly.

Strategic Asset Sampling. To operationalize Strategic Asset Sampling, the system follows a structured sequence to build visual reasoning and corresponding traces: starting with Task Selection, the pipeline pulls relevant elements from a pool of

10,000+ pre-rendered assets based on the selected task. These assets are then arranged into coherent visual configurations before moving to Trace Generation. We use LLMs to refine the trace text without viewing the assembled visuals, ensuring that linguistic diversity is enhanced while the alignment between logic and visual structure remains intact. The result of this coordinated process is the Compositional Puzzle Dataset, a robust, large-scale resource tailored for diagnostic evaluations of visual reasoning systems.

3.3 Foundatinoal Perception Dataset

The Perception Dataset comprises 30 foundational perceptual elements organized into a Cognitive Complexity Pyramid rather than a randomized collection, as shown in Figure 2. This systematic arrangement follows a rigorous logical progression from intrinsic object properties to quantitative analysis, effectively mapping the boundaries of visual intelligence across four distinct stages. This hierarchy is grounded in the functional architecture of the primate visual system (Ungerleider, 1982).

Figure 2: Cognitive Complexity Pyramid of the Perception Dataset, illustrating the four-stage logical progression from intrinsic geometric properties to advanced quantitative analysis.

Intrinsic Properties (Level 1). As the foundation of visual perception, this level evaluates a model's viewpoint invariance, the ability to maintain consistent object recognition across transformations such as rotation, scaling, and translation (DiCarlo and Cox, 2007). The evaluation focuses on core geometric features, such as symmetry and connectivity, that remain stable regardless of the viewing angle. We incorporate topological primitives (e.g., holes and contours), as these global traits are often processed prior to local details (Chen, 1982). Failure at this stage indicates that a model lacks a stable object concept, which is a prerequisite for any high-level reasoning.

Relational Properties (Level 2). Once individual objects are identified, this level shifts to spatial relationship logic. It tests the model's capacity to bind distinct entities into a coherent structure by assessing abstract states such as containment, tangency, and intersection (Kosslyn, 1987). These tasks evaluate the model's understanding of topological invariance, the ability to recognize relationships that remain constant even if the objects are stretched or distorted. This ensures the model captures fundamental spatial logic rather than relying on rigid coordinate matching or superficial patterns.

Global Syntax (Level 3). Mastery of global syntax requires a transition from local interactions to a unified scene understanding governed by coordinate systems and structural constraints. This stage aligns with the principle that global layout provides the necessary context for interpreting local details (Hochstein and Ahissar, 2002). By utilizing diverse grid architectures and path planning tasks, we evaluate scene grammar, the underlying rules that organize a visual environment (Võ and Wolfe, 2013). Models must operate within a global logic where every element is grounded by its specific coordinate and adherence to global rules.

Quantitative Analysis (Level 4). The highest cognitive level transitions from qualitative recognition to precise Quantitative Analysis. This stage focuses on numerical grounding—the ability to represent quantity independently of an object's specific appearance or scale (Nieder and Dehaene, 2009). It demands pixel-level parsing for the exact enumeration of points, lines, and areas. Because even a single-pixel deviation can lead to an incorrect result, this level precisely exposes a model's perceptual limits in bridging visual input with rigorous mathematical logic.

3.4 Compositional Puzzle Dataset

The Compositional Puzzle Dataset provides a high level extension that transforms atomic perception primitives into complex reasoning challenges. From the algorithmically synthesized pool, 20,000 samples are extracted to form the train set, while an additional 1,000 samples constitute the dedicated test set. These samples span 200 subclasses and require multi step induction or deduction to resolve the underlying logic. Each puzzle is paired with a verifiable reasoning trace that aligns intermediate logical steps with supporting visual evidence, providing a strong supervision signal for abstract graphic reasoning.

Reasoning Strategy Diversity. Analysis of the 20,000 reasoning traces reveals diverse logical patterns that can be grouped into 15 clusters. As

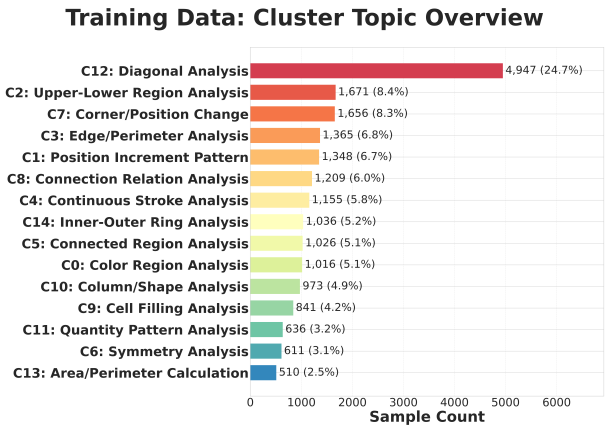


Figure 3: Statistical distribution of the 20,000 reasoning traces across 15 logical clusters.

shown in Figure 3, the dataset follows a long-tail distribution in which foundational strategies such as **Diagonal Analysis** account for 23.9% of samples. These common patterns are complemented by more specialized strategies (e.g., **Inner–Outer Ring Analysis** and **Quantity Pattern Analysis**), ensuring that models are exposed to both frequent and rare reasoning scenarios during training.

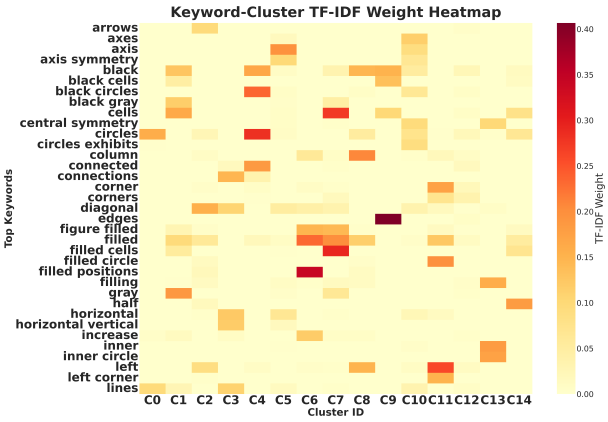


Figure 4: Keyword-signature heatmap demonstrating semantic grounding. Each reasoning cluster is associated with a distinct vocabulary, confirming that the generated traces are grounded in specific logical trajectories to support attributable evaluation.

Semantic Grounding and Attribution. We assess the semantic consistency of the dataset using keyword-signature analysis. As shown in Figure 4, the keyword cluster-weight heatmap indicates that each logic type is associated with a distinct vocabulary, therefore delineating different reasoning trajectories. By providing grounded yet linguistically diverse reasoning traces, the dataset supports attributable evaluation and enables precise diagnosis of whether failures arise from perceptual errors or incorrect rule application. This design maintains a training signal that is both logically consistent and linguistically diverse.

4 Experiments

In this section, we will introduce the implementation of our experiment. We evaluate the model performance in both visual reasoning benchmarks and foundational perception benchmarks.

4.1 Experimental Setup

Evaluated Models. We compare two open-source baselines, **Qwen3-VL-8B-Instruct** (Bai et al., 2025) and **InternVL3.5-VL-8B** (Wang et al., 2025), with their FLUSHPuzzle-tuned counterparts (**Qwen3-VL-FLUSHPuzzle** and **InternVL3.5-VL-FLUSHPuzzle**). The tuned models are optimized on FLUSHPuzzle using Group Relative Policy Optimization (GRPO) (Shao et al., 2024).

Implementation Details. We optimize Qwen3-VL-FLUSHPuzzle and InternVL3.5-VL-FLUSHPuzzle via a direct Reinforcement Learning (RL) framework on 20,000 samples, leveraging the GRPO algorithm to foster autonomous reasoning. The process employs the AdamW optimizer (1×10^{-6} learning rate) with a global batch size of 192 distributed across 96 NVIDIA H800 GPUs. Each prompt involves a rollout of 8 responses, with policy stability maintained by a low-variance KL-divergence estimator ($\beta = 0.01$). While the training sequence length is capped at 4,096 tokens, inference extends to 16,384 (Temp 0.2, Top-p 0.8) to capture exhaustive reasoning traces. The models are trained for 4 epochs to ensure the convergence of complex logical strategies.

4.2 Evaluation of Foundational Perception

To provide a rigorous assessment of visual capabilities, aside from the public benchmarks and FLUSHPuzzle-test, we introduce a Unified Perception Evaluation Framework that integrates axiomatic geometric probes with the BLINK benchmark. The framework evaluates models across seven dimensions, ranging from atomic shape recognition to broader visual grounding in diverse contexts. By combining six specialized tasks with BLINK, we assess both foundational geometric primitives and the ability to generalize to external, high-quality visual data. To ensure fairness and establish a strong reference point, we evaluate recent state-of-the-art public models on the six carefully constructed tasks. The results are shown in Table 1.

Shape Identification Proficiency (T1). This dimension measures a model's ability to recognize

Model	Foundational Perception Tasks (Accuracy, %)						BLINK (Accuracy, %)
	T1	T2	T3	T4	T5	T6	
General and Commercial Models							
Gemini 3 Pro Preview	**91.67**	**96.33**	**99.44**	**95.01**	**96.56**	**97.80**	**79.80**
Qwen3-VL-32B-Instruct	87.41	77.67	94.26	31.21	75.56	85.80	67.30
Doubao-1.5-VL	78.89	91.33	96.85	58.34	90.22	92.80	72.10
ERNIE-4.5-VL-A3B-Thinking	74.44	85.00	83.89	26.77	67.78	90.40	76.33
Baseline vs. Qwen3-VL-FLUSHPuzzle							
Qwen3-VL-8B (Baseline)	**86.85**	55.67	83.52	**33.02**	70.56	70.20	66.91
Qwen3-VL-FLUSHPuzzle	85.37	**66.33**	**91.11**	32.43	**74.89**	**71.20**	**67.07**

Table 1: Performance across the Unified Perception Framework. T1-T6 represent Shape Recognition, Size Comparison, Difference Detection, Angle Recognition, Intersection Counting, and Length Perception, respectively. For specialized models, **bold** indicates the better result between the baseline and our version.

geometric categories across both basic shapes and visually similar, easily confusable shapes. It distinguishes coarse category recognition from the fine-grained discrimination required for near-identical structures. Many proprietary models achieve high accuracy on basic shapes but degrade when evaluated on similar shapes. In contrast, Qwen3-VL-FLUSHPuzzle remains stable on this dimension, suggesting that the model preserves its fundamental geometric classification capabilities despite undergoing rigorous reasoning-focused supervision.

Relative Magnitude Perception (T2). This task systematically assesses the model's ability to perceive relative spatial scale and hierarchical ordering within a constrained, fixed canvas. Models are required to precisely identify the extremum element—either the largest or smallest—among multiple co-occurring instances of the same shape, with task difficulty finely controlled by varying the incremental size differences. The empirical results expose a substantial performance gap between 8B-parameter baselines and significantly larger proprietary models. Notably, targeted training on FLUSHPuzzle elevates accuracy from 55.67% to 66.33%, strongly suggesting that relative-magnitude perception inherently benefits from structured geometric supervision.

Anomaly Localization Performance (T3). This task evaluates anomaly detection in a 2×2 grid containing one deviant shape among three identical items. It tests visual discrimination across both basic shapes and visually similar variants. Most strong models perform well on this dimension, with Gemini 3 Pro achieving near-perfect accuracy. Compared with the baseline, Qwen3-VL-FLUSHPuzzle improves 7.59%, indicating that learning to isolate geometric anomalies is an important prerequisite for solving abstract visual puzzles.

Angle Quantification Precision (T4). This task evaluates angle recognition over the range $10°$–$350°$ using 1,705 samples. A prediction is counted as correct if the absolute error is within $10°$. The results indicate that angle perception remains a major bottleneck for most MLLMs, with performance strongly affected by reflex-angle cases. Larger models such as Gemini 3 Pro achieve higher precision, whereas smaller models and some proprietary systems exhibit substantially higher error rates. Qwen3-VL-FLUSHPuzzle remains comparable to the baseline, suggesting that angle quantification still requires further perceptual improvement.

Topological Parsing Accuracy (T5). We measure topological parsing by counting intersections in complex line-segment configurations. The task requires models to disentangle overlapping structures and accurately enumerate contact points. Gemini 3 Pro and Doubao-1.5VL perform strongly on this benchmark. Qwen3-VL-FLUSHPuzzle improves by 4.33 percentage points (from 70.56% to 74.89%), suggesting that FLUSHPuzzle training helps models resolve complex spatial intersections.

Metric Grounding Capability (T6). This task assesses length perception. We count predictions as correct when they fall within 5 mm of the ground truth. Gemini 3 Pro achieves near-perfect performance under this criterion, and smaller models show reasonable proficiency. Training with FLUSHPuzzle yields a $\sim$1 % gain for Qwen3-VL-FLUSHPuzzle, suggesting that supervision with precise metric measurements can strengthen grounding in physical and diagrammatic contexts.

External Perception Evaluation To test whether gains on the axiomatic probes generalize beyond synthetic settings, we also evaluate models on the public BLINK benchmark (Fu et al., 2024). BLINK measures core visual skills such as depth estima-

tion and spatial orientation across diverse scenarios that discourage linguistic shortcuts. Qwen3-VL-8B scores 66.91%, while Qwen3-VL-FLUSHPuzzle scores 67.07%. Although the improvement is not significant, it suggests that geometric axioms and grounded reasoning traces can modestly enhance visual robustness beyond puzzle solving.

4.3 Evaluation of Graphic Puzzle Reasoning

To provide a comprehensive assessment, we evaluate models on our FLUSHPuzzle-Test set together with four public benchmarks: LogicVista (Xiao et al., 2024), VisuLogic (Xu et al., 2025), VisualPuzzles (Song et al., 2025), and VisualSphinx V1 (Feng et al., 2025). These datasets require multi-step induction, topological reasoning, and abstract pattern recognition. Overall, the results show that FLUSHPuzzle training substantially reduces the performance gap between small open-source models and strong proprietary systems.

Reasoning Performance on FLUSHPuzzle-Test. As shown in Table 2, both Qwen3-VL-FLUSHPuzzle and InternVL3.5-FLUSHPuzzle improve markedly on FLUSHPuzzle-Test. Qwen3-VL-FLUSHPuzzle achieves 43.00% accuracy, a 15.80% increase over baseline. InternVL3.5-FLUSHPuzzle reaches 41.23%, improving by 14.63% over baseline. These 8B-parameter models are competitive with Gemini 3 Pro (45.00%) and outperform Doubao-1.5VL (34.40%). The gains suggest that structured reasoning traces and axiomatic perceptual grounding enable smaller models to solve complex logic puzzles that previously required larger proprietary systems.

Generalization to Public Benchmarks. Both tuned models generalize well across public reasoning benchmarks. On VisualSphinx V1, Qwen3-VL-FLUSHPuzzle improves from 40.32% to 52.41% (+12.09 percentage points), while InternVL3.5-FLUSHPuzzle increases from 36.36% to 49.92% (+13.56 percentage points). Consistent gains are also observed on LogicVista, VisuLogic, and VisualPuzzles for both models, with InternVL3.5-FLUSHPuzzle showing particularly strong transfer on VisuLogic (+2.38 percentage points) and VisualPuzzles (+4.69 percentage points). Together, these results indicate that the reasoning patterns learned from FLUSHPuzzle transfer to diverse abstract graphic tasks, thereby strengthening the reasoning capability of different multimodal.

4.4 Analysis: Does Grounded Training Elicit Sound Logical Methodologies or Heuristic Shortcuts?

From Table 2, we see the performance boost when models are grounded-trained on FLUSHPuzzle. To evaluate whether gains stem from authentic structural reasoning or statistical shortcuts, we conduct a diagnostic analysis by clustering the prediction patterns of both the Baseline and Qwen3-VL-FLUSHPuzzle.

From Strategy Collapse to Strategy Diversification. As shown in Figure 5, the baseline model exhibits a "heuristic collapse" phenomenon, with an overwhelming 76.0% of its predictions concentrated in Cluster 11 (Quantity Pattern Analysis). This skewed distribution signals a pathological over-reliance on element counting as a universal shortcut for reasoning tasks, rather than adapting to diverse problem structures. In stark contrast, grounded training effectively mitigates this bias: it reduces the proportion of predictions in C11 to 64.7% while redistributing outputs across specialized clusters, including C0, C4, and C13. This notable shift underscores the first key contribution of grounded training: it expands the model's strategic reasoning space, curbing the tendency to default to a single magnitude-based heuristic and enabling more flexible, task-aligned problem-solving.

Finding 1

Grounded training helps break the model's reliance on a single heuristic strategy, enabling a transition from strategy collapse to a more diverse reasoning repertoire.

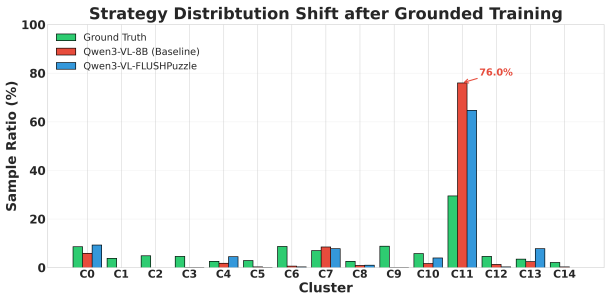


Figure 5: Prediction Distribution Shift across 15 Strategy Clusters.

Functional Recovery of Logical Consistency. While Finding 1 delineates the diversity of strategies employed, Figures 6 and 7 empirically validate the effectiveness of these strategic choices. The baseline model reallocates 90% of C2 (Upper-Lower Region) and 87% of C7 (Symmetry) sam-

Model	FLUSHPuzzle-Test	LogicVista	VisuLogic	VisualPuzzles	VisualSphinx
General and Commercial Models					
Gemini 3 Pro	45.00	80.80	37.60	71.48	-
Doubao-1.5VL	34.40	74.55	35.30	54.88	-
Baseline vs. FLUSHPuzzle					
Qwen3-VL-8B-Instruct (Baseline)	27.20	57.14	20.70	37.33	40.32
Qwen3-VL-FLUSHPuzzle	**43.00**	**58.93**	**25.50**	**41.10**	**52.41**
InternVL3.5-8B (Baseline)	26.60	52.12	31.10	35.96	36.36
InternVL3.5-FLUSHPuzzle	**41.23**	**54.60**	**33.48**	**40.65**	**49.92**

Table 2: Performance comparison on graphic puzzle reasoning benchmarks (Accuracy, %). **FLUSHPuzzle-Test** refers to our specialized evaluation set. For specialized models, **bold** indicates the superior performance achieved.

ples to quantity-based heuristics, indicating a failure to recognize task-specific geometric rules. In contrast, Qwen3-VL-FLUSHPuzzle achieves notable diagonal pattern recovery across C0 (64%), C5 (48%), and C13 (60%). These results demonstrate that the model does not merely randomly adopt novel strategies; instead, it has learned to systematically map explicit structural constraints to their corresponding visual features.

Finding 2

The model demonstrates a measurable recovery of logical methodology consistency, proving it can decouple specific geometric rules from global counting correlations.

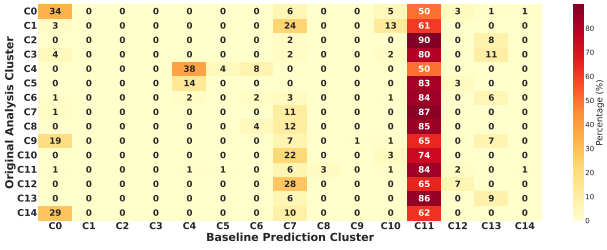


Figure 6: Strategy Confusion Matrix of Baseline Model.

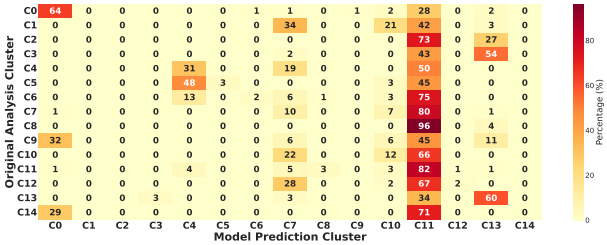


Figure 7: Strategy Confusion Matrix of Qwen3-VL-FLUSHPuzzle.

The Persistence of the Path of Least Resistance. Despite these improvements, the quantity-based heuristic remains a "path of least resistance" for high-complexity tasks. This is evidenced by persistent high migration rates in clusters such as C8 (96% to C11) and C2 (73% to C11). While grounded training successfully elicits authentic logic in structural tasks, magnitude-based shortcuts still act as a latent fallback when geometric patterns are less salient. Our model represents a "reasoner-in-transition"—capable of authentic structural alignment but still susceptible to lingering statistical traps.

Finding 3

Grounded training initiates the shift toward sound logical methodologies, yet the quantity shortcut remains a persistent fallback for the most abstract and complex clusters.

5 Conclusion

We address systemic MLLM reasoning deficiencies via FLUSH-Gen, a rule-based framework that decouples perception from logic using hierarchical primitives and verifiable reasoning traces. Our benchmark, FLUSHPuzzle, enables fine-grained diagnosis, demonstrating that targeted geometric supervision empowers 8B models to rival commercial giants. We liken the puzzle-solving process to identifying hand patterns in poker; success requires the model to look beyond raw element counts to discern the governing logical "hand types" currently in play. Crucially, our analysis reveals that while MLLMs are prone to "heuristic collapse" toward the counting strategy, grounded training effectively alleviates this homogenization by eliciting authentic structural logic. Although magnitude-based shortcuts remain a persistent "path of least resistance" in high-complexity scenarios, this work proves that structural alignment is essential for transitioning models from statistical pattern matching to robust reasoning. Preserving strategic diversity against persistent heuristic traps remains a primary challenge for future research.

6 Limitations

Despite its contributions, FLUSHPuzzle has several constraints. While grounded training significantly mitigates strategy homogenization, a residual bias toward quantity-based heuristics persists as a "path of least resistance" for high-complexity tasks, indicating that the transition from statistical shortcuts to authentic structural logic remains incomplete. Additionally, the benchmark's current restriction to abstract graphic domains leaves a gap between axiomatic geometric reasoning and the complex semantic understanding required for natural scenes. Finally, the framework's focus on single-image paradigms suggests a need to extend hierarchical reasoning traces to long-context visual sequences and broader multi-modal interactions in future research.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Lin Chen. 1982. Topological structure in visual perception. *Science*, 218(4573):699–700.

James J DiCarlo and David D Cox. 2007. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341.

Yichen Feng, Zhangchen Xu, Fengqing Jiang, Yuetai Li, Bhaskar Ramasubramanian, Luyao Niu, Bill Yuchen Lin, and Radha Poovendran. 2025. Visualsphinx: Large-scale synthetic vision logic puzzles for rl. *arXiv preprint arXiv:2505.23977*.

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

Shaul Hochstein and Merav Ahissar. 2002. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5):791–804.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Stephen M Kosslyn. 1987. Seeing and imagining in the cerebral hemispheres: a computational approach. *Psychological review*, 94(2):148.

Jinhao Li, Zijian Chen, Lirong Deng, Changbo Wang, and Guangtao Zhai. 2025. Mmreid-bench: Unleashing the power of mllms for effective and versatile person re-identification. *arXiv e-prints*, pages arXiv–2508.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

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

Andreas Nieder and Stanislas Dehaene. 2009. Representation of number in the brain. *Annual review of neuroscience*, 32(1):185–208.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, and 1 others. 2025a. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26147–26159.

Kexian Tang, Junyao Gao, Yanhong Zeng, Haodong Duan, Yanan Sun, Zhening Xing, Wenran Liu, Kaifeng Lyu, and Kai Chen. 2025b. Lego-puzzles: How good are mllms at multi-step spatial reasoning? *arXiv preprint arXiv:2503.19990*.

Leslie G Ungerleider. 1982. Two cortical visual systems. *Analysis of visual behavior*, 549:chapter–18.

Melissa L-H Võ and Jeremy M Wolfe. 2013. Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychological science*, 24(9):1816–1823.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Anjiang Wei, Yuheng Wu, Yingjia Wan, Tarun Suresh, Huanmi Tan, Zhanke Zhou, Sanmi Koyejo, Ke Wang, and Alex Aiken. 2025. Satbench: Benchmarking llms' logical reasoning via automated puzzle generation from sat formulas. *arXiv preprint arXiv:2505.14615*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, and 1 others. 2025. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. 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.

A Semantic Clustering Methodology for Reasoning Traces

A.1 Data Preprocessing and Feature Extraction

To ensure the clustering process focuses on the underlying logical patterns, we apply a rigorous preprocessing pipeline to the training dataset consisting of 20,000 samples. This procedure isolates structural reasoning from superficial classification results to provide a clean signal for semantic grouping.

Semantic Logic Isolation We isolate the reasoning logic from the assistant responses by extracting only the Analysis section while discarding the single-letter Answer labels. This ensures the clustering algorithm is driven by linguistic reasoning structures rather than classification outcomes. The Analysis section serves as the most representative component of the problem-solving pattern within each puzzle.

Reasoning Extraction Protocol The extraction is performed via regular expression matching, targeting the text segment between the Analysis header and the Answer marker. If no such markers are found, the system defaults to the full content to preserve data integrity across the 20,000 samples. This standardized format allows the subsequent vectorization stage to operate on purely logical text.

A.2 Text Vectorization and Numerical Representation

The extracted reasoning text is converted in high-dimensionalnal vectors using the Term Frequency Inverse Document Frequency (TF-IDF) method to capture the importance of specific geometric and logical terms.

TF-IDF Framework The vectorization process utilizes a vocabulary of the top 5000 features, incorporating both unigrams and bigrams to capture complex phrases such as axis symmetry or black cells. We implement a minimum document frequency of 5 to eliminate rare noise and a maximum frequency of 0.8 to remove overly common terms that lack discriminative power. English stop words are removed to focus exclusively domain-specific semantic content.

Mathematical Weighting Strategy For each term t in document d, the weight is calculated by multiplying the local term frequency by the global inverse document frequency. The formula for this representation is:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t) \tag{1}$$

The Inverse Document Frequency (IDF) is defined by the total number of documents N relative to the count of documents containing the specific term:

$$\text{IDF}(t) = \log \frac{N}{|\{d : t \in d\}|} \tag{2}$$

A.3 Clustering Implementation and Statistical Evaluation

We apply the K-Means algorithm to the resulting sparse matrix to group semantically similar reasoning traces into discrete logical categories.

K-Means Execution and Optimal K Selection The algorithm uses Euclidean distance in the feature space to assign each sample to the nearest centroid vector. The distance between a sample x and a centroid c is calculated as:

$$d(x, c) = \sqrt{\sum_{i=1}^{5000} (x_i - c_i)^2} \tag{3}$$

The optimal number of clusters, $K = 15$, was determined by evaluating the within-cluster sum of squares through the elbow method. This selection balances the need for sufficient granularity to capture distinct reasoning patterns against the risk of over-fragmentation.

Quantitative Distribution and Semantic Attribution The resulting distribution reveals 15 distinct clusters characterized by unique keyword signatures extracted from the centroids, as shown in Figure 8. The largest group, Cluster C12, accounts for 24.7% of the data and focuses on Diagonal Analysis involving corner and region patterns. Specialized clusters include C8 for Connection Relation Analysis, C11 for Quantity Pattern Analysis, and C10 for Column and Shape Analysis. Each sample in the final dataset is appended with a cluster field ranging from 0 to 14 to facilitate targeted evaluation based on these semantically grounded reasoning categories.

Rank	ID	Topic	Samples	Ratio	Top Keywords
1	C12	Diagonal Analysis	4,947	24.7%	corners, regions, pattern, positions
2	C2	Upper-Lower Region Analysis	1,671	8.4%	diagonal, arrows, right, left
3	C7	Corner/Position Change	1,656	8.3%	filled cells, cells, filled, figure filled
4	C3	Edge/Perimeter Analysis	1,365	6.8%	connections, horizontal, vertical, horizontal vertical
5	C1	Position Increment Pattern	1,348	6.7%	gray, cells, wavy, black
6	C8	Connection Relation Analysis	1,209	6.0%	column, shapes, right, left
7	C4	Continuous Stroke Analysis	1,155	5.8%	circles, black circles, connected, black
8	C14	Inner-Outer Ring Analysis	1,036	5.2%	upper, lower, half, upper lower
9	C5	Connected Region Analysis	1,026	5.1%	symmetric, axis, symmetry, axis symmetry
10	C0	Color Region Analysis	1,016	5.1%	circles, tangent, straight, lines
11	C10	Column/Shape Analysis	973	4.9%	symmetry, axes, central symmetry, axis
12	C9	Cell Filling Analysis	841	4.2%	edges, perimeter, black, shaded
13	C11	Quantity Pattern Analysis	636	3.2%	left, filled circle, position, corner
14	C6	Symmetry Analysis	611	3.1%	filled positions, positions, filled, figure filled
15	C13	Area/Perimeter Calculation	510	2.5%	outer ring, inner, ring, inner circle

Figure 8: Semantic taxonomy and keyword-based attribution of the reasoning clusters.

B Experiments

B.1 Training Dynamics and Checkpoint Analysis

To evaluate the efficacy of GRPO and determine the optimal training duration, we conduct a step-wise ablation analysis across six distinct visual reasoning benchmarks. This analysis tracks the model performance from the baseline (Qwen3-VL-8B-Instruct) to the final checkpoint at Step 416 (4 epoch).

Performance Surge and Benchmarking Results As illustrated in Figure 9, the training progress reveals a non saturating upward trajectory for complex reasoning tasks while maintaining robust stability on foundational perception benchmarks. The quantitative results, including the average accuracy across all tasks, are summarized in Table 3.

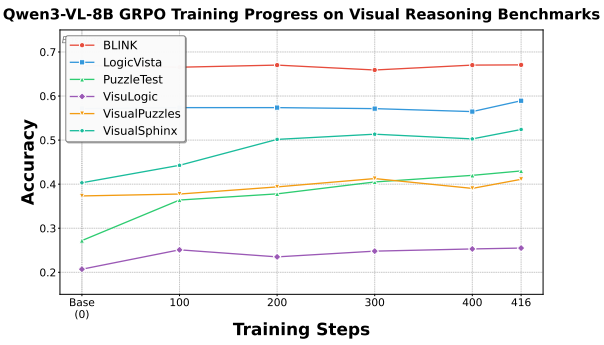


Figure 9: Migration paths of reasoning-strategy .

Reasoning Intensive Surges Benchmarks requiring intensive visual reasoning exhibit a profound performance surge, with PuzzleTest ascending from 27.2% to 43.0%, representing a 58.09% relative gain. Similarly, VisualSphinx and VisuLogic show steady improvements of 29.99% and 23.19% respectively, indicating that the model is successfully internalizing structured logic. The consistent upward trend in these tasks suggests that GRPO effectively elicits specialized problem solving capabilities without reaching a saturation plateau within the observed steps.

Perceptual Robustness and Knowledge Retention In contrast to the reasoning gains, benchmarks focused on foundational perception and standard logic, such as BLINK and LogicVista, remain remarkably stable. BLINK maintains its performance with a negligible change of 0.24%, while LogicVista sees a modest 3.13% improvement. This stability demonstrates that the training process successfully mitigates catastrophic forgetting, preserving the foundational instruction following and perceptual capabilities of the base model while specifically augmenting its reasoning depth.

Fluctuation and Overfitting Risks Performance on VisualPuzzles peaked at Step 300 before exhibiting slight volatility, suggesting the potential onset of domain specific overfitting in the later stages of training. Despite a minor average dip at Step 400, the model recovered at Step 416 to reach its highest overall accuracy of 48%. This lack of a clear performance ceiling implies that further training iterations could yield additional gains for non saturated benchmarks.

Optimal Checkpoint Determination We select Step 416 as the optimal model for final evaluation, as it achieves the global maximum in average performance and attains the best or near best accuracy

Checkpoint	BLINK	LogicVista	PuzzleTest	VisuLogic	VisualPuzzles	VisualSphinx	Average
Baseline Model							
Base (0)	66.91	57.14	27.20	20.70	37.33	40.32	41.60
GRPO Training Steps							
Step 100	66.54	57.37	36.40	25.10	37.76	44.28	44.58
Step 200	67.02	57.37	37.80	23.50	39.38	50.16	45.87
Step 300	65.91	57.14	40.50	24.80	**41.27**	51.34	46.83
Step 400	67.02	56.47	42.00	25.30	39.04	50.27	46.68
Step 416	**67.07**	**58.93**	**43.00**	**25.50**	41.10	**52.41**	**48.00**

Table 3: Training dynamics of Qwen3-VL-8B during GRPO optimization (Accuracy, %). **Step 416** is selected as the optimal checkpoint for exhibiting superior performance in 5 out of 6 benchmarks. **Bold** indicates the highest accuracy achieved per category.

on five out of six benchmarks. This checkpoint represents the most balanced configuration of enhanced logical depth and retained perceptual sensitivity achieved during the GRPO process.

B.2 Comparative Analysis: GRPO vs. Supervised Fine-tuning

To further validate the efficiency of Group Relative Policy Optimization (GRPO) in eliciting reasoning capabilities, we conduct a comparative study against traditional Supervised Fine Tuning (SFT) methods. This comparison investigates the trade offs between in domain specialization and out of domain generalization across two SFT configurations: one with a frozen vision tower and one with an unfrozen vision tower. Both SFT models were trained on 20,000 samples from the FLUSHPuzzle dataset for 4 epochs, with the best performing checkpoints selected to ensure a fair and rigorous comparison against the GRPO results as shown in Table 4 and Figure 10.

In Domain Specialization and Fitting Efficiency The results demonstrate that SFT is highly efficient at fitting the training data distribution, particularly in the PuzzleTest benchmark where the unfrozen SFT model achieves a peak accuracy of 64.00%. This represents a 136% improvement over the base model, significantly higher than the 58% gain achieved by GRPO. However, this high performance is primarily localized to the training domain, suggesting that while the FLUSHPuzzle dataset possesses high pedagogical value for supervised learning, the SFT method prioritizes pattern matching over broad logical transfer.

Generalization Deficits and Catastrophic Forgetting Despite the in-domain success of SFT, it exhibits severe generalization deficits on out-of-domain benchmarks. While GRPO achieves an average performance gain of 14 percent across general reasoning tasks, SFT Unfreeze suffers a significant degradation of 10 percent, falling below the baseline in several categories such as LogicVista and VisualPuzzles. This indicates that a narrow focus on a single logic type during SFT leads to catastrophic forgetting. Interestingly, freezing the vision tower during SFT mitigates this loss to some extent, but the overall average performance remains inferior to the GRPO paradigm.

Efficacy of Data and Training Paradigms Our findings suggest that the performance degradation observed in SFT is not a result of poor data quality but rather a consequence of the training paradigm. The FLUSHPuzzle data is clearly suitable for supervised learning given its high in-domain scores; however, the narrowness of the task distribution induces a strategy collapse. To maximize the utility of such specialized reasoning data without compromising generalization, it should be utilized within a mixed training or multi task framework alongside general vision language datasets. Compared to SFT, GRPO serves as a more robust elicitation method, maintaining perceptual stability while successfully augmenting the model underlying reasoning depth.

C Prompt

The FLUSH-Gen utilizes specialized prompt templates to ensure high linguistic quality and logical consistency across its synthesized reasoning traces. As shown in Table 5 and Table 6, these prompts are applied during the Blind LLM Polishing and Reasoning Traces generation stages, where an LLM refines symbolic metadata into natural language

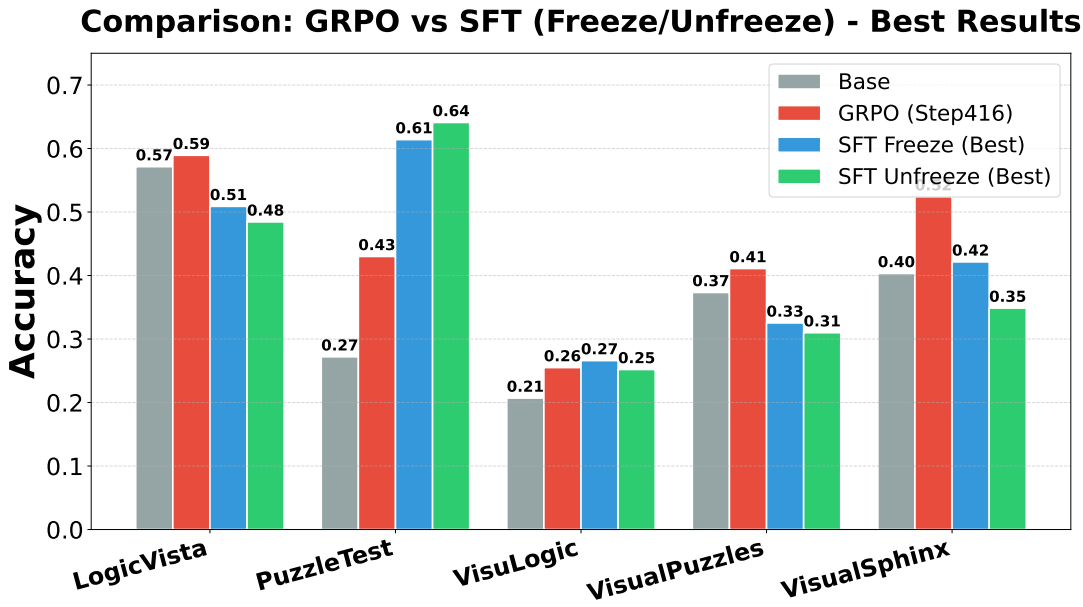


Figure 10: Performance comparison of GRPO and SFT variants across visual reasoning benchmarks. While SFT-based methods exhibit superior performance on the domain-specific PuzzleTest, GRPO demonstrates more robust generalization across out-of-distribution benchmarks, effectively mitigating the catastrophic forgetting observed in SFT models.

without direct visual access to the images, thereby preventing visual hallucination.

D Visual Perception Benchmark

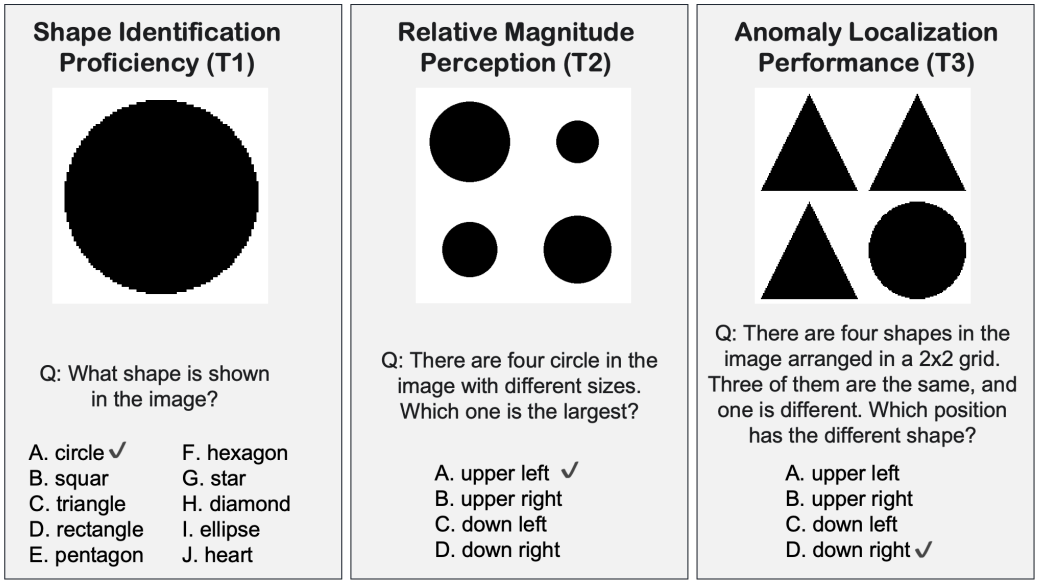
To provide concrete visualization of the perception benchmark dimensions, Figure 11 presents representative samples for all six tasks. The visualization is divided into two phases: Panel (a) illustrates the discriminative tasks (T1–T3), covering Shape Identification, Relative Magnitude Perception, and Anomaly Localization, which assess the model's fundamental ability to distinguish geometric categories and spatial hierarchies. Panel (b) details the quantitative and topological tasks (T4–T6)specifically Angle Quantification, Topological Parsing, and Metric Grounding—which require the model to perform precise geometric measurements and disentangle complex structural intersections.

E Visual Perception Examples

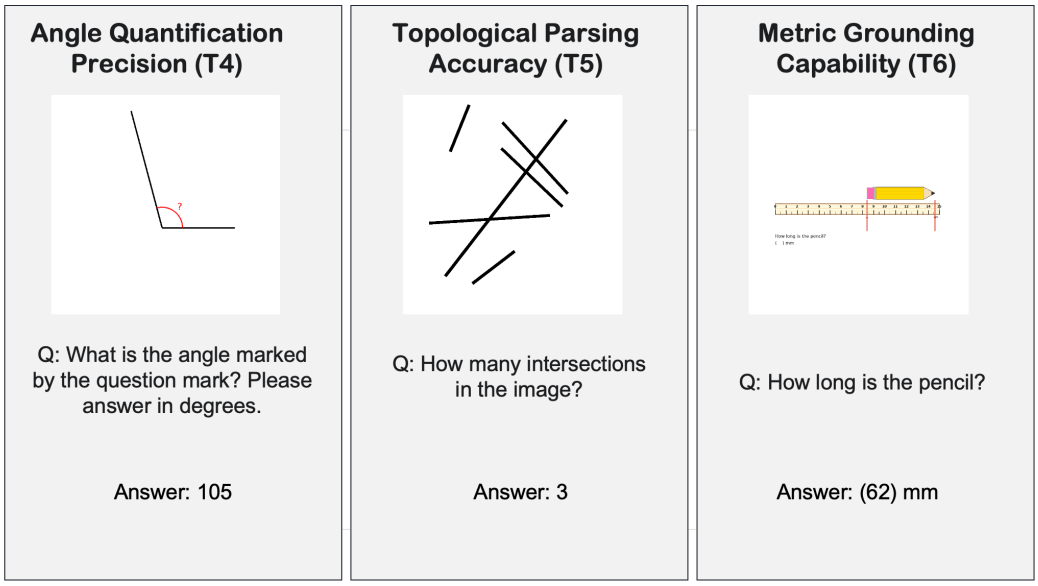
To provide a clear understanding of FLUSHPuzzle in our work, we present a selection of 5 samples from our Perception Dataset in Table 7.

F Visual Reasoning Examples

This appendix provides a comprehensive collection of detailed examples from Table 8 to Table 22, illustrating the structural and logical diversity inherent within the FLUSHPuzzle dataset. These cases serve as qualitative studies derived from the Algorithmic Logic Synthesis module of FLUSH-Gen, demonstrating how abstract symbolic rules are deterministically mapped to specific visual layouts, such as grid based or spatial configurations.



(a) Examples of T1–T3



(b) Examples of T4–T6

Figure 11: Samples of tasks from T1 to T6. (a) Illustrates the T1-T3, while (b) shows T4-T6s.

Benchmark	Base	GRPO	SFT Freeze	SFT Unfreeze
In Domain Performance				
PuzzleTest	27.20	43.00	61.00	**64.00**
Generalization Benchmarks				
LogicVista	57.14	**58.93**	51.00	48.00
VisuLogic	20.70	25.50	**27.00**	25.00
VisualPuzzles	37.33	**41.10**	33.00	31.00
VisualSphinx	40.32	**52.41**	42.00	35.00
Statistical Summary				
Average Accuracy	37.00	**44.19**	42.80	40.60

Table 4: Comparative performance across different training paradigms (Accuracy, %). **GRPO** demonstrates superior generalization across diverse reasoning tasks, whereas **SFT** variants exhibit extreme in domain fitting at the expense of external reasoning stability. **Bold** indicates the highest accuracy per category.

Prompt: Image Caption Polishing

You are an expert in refining image descriptions to make them clear, natural, and professional.

Please polish the following English image description while:

1. Maintaining all factual information and details
2. Improving clarity and readability
3. Using natural, idiomatic English
4. Keeping technical terms accurate
5. Ensuring smooth flow and professional tone

Original Description:
{text}

Output Requirements:
Return ONLY the polished version, without explanations.

Table 5: The original prompt template used for refining and professionalizing raw image descriptions.

Prompt: Deductive Reasoning Rephrasing

You are an expert in graphical reasoning and English linguistic polishing. Although you cannot see the actual images, you are provided with the question text, analysis, options, and the correct answer.

Task: Rephrase the question and the analysis in English. Do not change the underlying reasoning logic, the identified patterns, or the final answer.

[Question Rephrasing]
- Rewrite the question using more natural and fluent English.
- Maintain the original intent: "Choose the fifth image that best fits the pattern established by the first four."
- Do not copy the original text verbatim; there must be noticeable changes in word choice or sentence structure.
- At the end of the question, add a concluding sentence prompting the user to "make the correct choice based on the summarized pattern" (you may use your own phrasing).

[Analysis Rephrasing] Reorganize the original analysis into a clear, structured format:
1. Summarize the key features of the first four images (based on the provided description).
2. Explain the underlying logic or pattern observed across the first four images.
3. Infer the expected characteristics of the fifth image.
4. Identify the matching correct option: {answer}.

[Prohibited Content]
- Do not copy sentences directly from the original question or analysis.
- Do not add new reasoning or patterns that are not present in the original text.
- Do not change the correct answer.

[Output Format]
Please output strictly in the following JSON format:
```
{{ "rephrased_question": "...", "rephrased_analysis": "..." }}
```

Prompt: Inductive Reasoning Rephrasing

You are an expert in graphical reasoning and English linguistic polishing. Although you cannot see the actual images, you are provided with the question text, analysis, options, and the correct answer.

Your task is to rewrite the question and analysis to make them more fluent and logically structured, without changing the original intent or reasoning logic.

[Question Rewriting Requirements]
- Use entirely new phrasing to rewrite the original question; do not simply repeat or copy the original sentences.
- Change multiple words or adjust the syntax to ensure the text is significantly different from the original.
- Ensure the core meaning remains the same: Six images need to be divided into two groups of three based on a shared rule.
- Do not reveal the grouping rule prematurely in the question.
- At the end of the question, add a prompt indicating that the user "needs to make the correct selection based on the grouping pattern of the figures".

[Analysis Rewriting Requirements] Reorganize the original analysis into a clearer structure, including the following points:
1. Provide an overview of the main features of the six images, either individually or in clusters (based on the original analysis).
2. Point out the core rule or commonality used for grouping.
3. Clearly state which three images belong to the first group and which three belong to the second group.
4. Ensure the conclusion remains consistent: The correct answer is {answer}.

[Prohibited Content]
- Do not copy sentences directly from the original question or analysis.
- Do not add non-existent reasoning or invent incorrect rules.
- Do not change the correct answer.

[Output Format]
Please strictly output the following JSON structure without any additional explanation:
```
{{ "rephrased_question": "...", "rephrased_analysis": "..." }}
```

Table 6: English prompt templates for deductive and inductive reasoning task rephrasing.

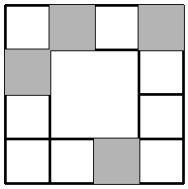
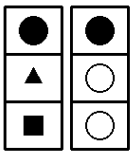
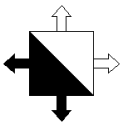
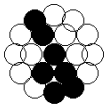
Sample ID & Image	Detailed Image Caption (Ground Truth Description)
 **ID0**	A 4×4 grid with a removed 2×2 center, forming a hollow loop of 12 cells. Among these, 4 cells are shaded gray at (Row 2, Column 1), (Row 1, Column 4), (Row 4, Column 3), and (Row 1, Column 2). A clockwise overview indicates Row 1 is filled at Columns 2/4, Row 2 at Column 1, and Row 4 at Column 3.
 **ID1**	Features a dual-column layout (3×2) with six positions. The left column is fully occupied by a black circle, triangle, and square. The right column contains only one black circle at Position 0, followed by two empty circular outlines. Cell dimensions approx. 60×50 with a spacing of 10.
 **ID2**	A central square frame with four outward-pointing arrows. The interior is bisected by a main diagonal (top-left to bottom-right). Shading is applied to the left and bottom arrows, while only the lower triangular region inside the square is filled.
 **ID3**	A hexagonal cluster of 19 circles (7 black, 12 white). The black-shaded region constitutes an unbroken, single-stroke sequence that bisects the white circles into two separate, non-connected regions.
 **ID4**	A 4×4 grid with 3 black-shaded cells at (R1, C1), (R1, C2), and (R2, C1). The remaining 13 cells are white. The cumulative perimeter of the shaded region is exactly 8 units.

Table 7: Examples of Foundational Perception Dataset of FLUSHPuzzles

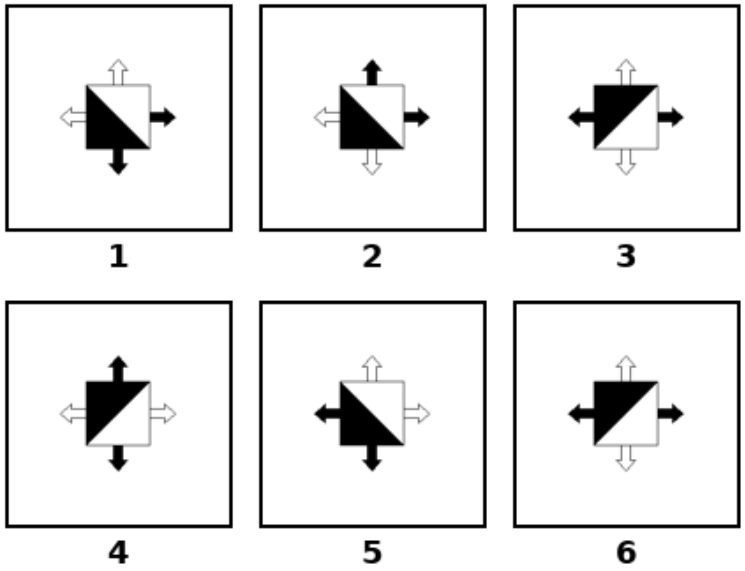
Case Study: Cluster 0 — Geometric Diagonal Reasoning

Question

Divide the six images into two groups of three based on a common pattern. Identify the numbers for each group and select the correct option representing the grouping rule.

Options:

A. {5, 2, 6}, {4, 3, 1} **B.** {5, 2, 4}, {3, 6, 1} **C.** {4, 5, 1}, {6, 3, 2} **D.** {5, 1, 2}, {6, 4, 3}



Logical Analysis

As illustrated in figure, the grouping logic is derived from the geometric symmetry of the filled regions and the diagonal division axes:

- **Group I (Figures 1, 2, 5):** Lower-left triangular region is filled via main diagonal division.
- **Group II (Figures 3, 4, 6):** Upper-left triangular region is filled via anti-diagonal division.

Correct Answer: D

Table 8: Overview of the Geometric Diagonal Reasoning Case Study.

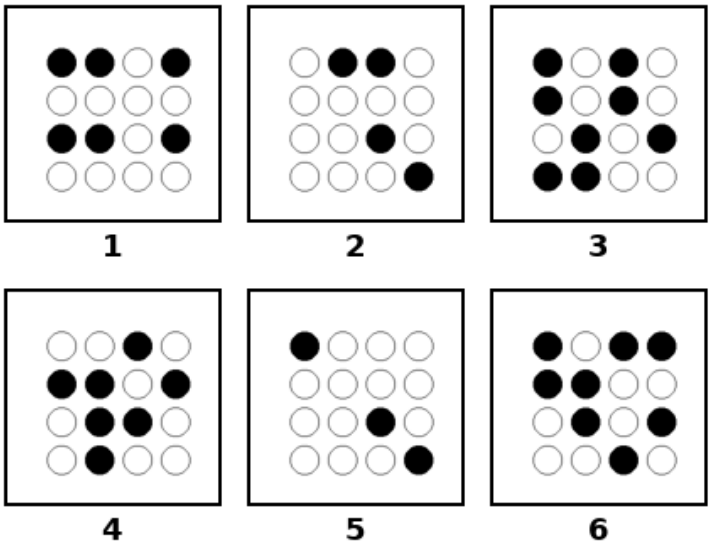
Case Study: Cluster 1 — White Region Connectivity Pattern

Question

Based on the pattern of the figures, divide the six images into two groups, each containing three images, and determine the grouping. Choose the correct option according to the grouping pattern.

Options:

A. {4, 1, 6}, {3, 5, 2} **B.** {3, 5, 1}, {6, 4, 2} **C.** {3, 6, 4}, {2, 5, 1} **D.** {5, 4, 6}, {3, 2, 1}



Logical Analysis

As illustrated in figure, each figure is a 4×4 circular grid. The core reasoning pattern is determined by the number of connected regions formed by the white (unfilled) circles through adjacency:

- **Group I (Figures 1, 2, 5):** In these figures, all white circles are mutually connected, forming exactly **one connected region**.
- **Group II (Figures 3, 4, 6):** In these figures, the placement of black circles partitions the white circles into **multiple separate connected regions**.
  - Fig 3: 4 regions; Fig 4: 5 regions; Fig 6: 4 regions.

Connectivity serves as the fundamental topological feature for this classification, abstracting away the raw count of black elements. This pattern aligns with the classification in Option C.

Correct Answer: C

Table 9: Overview of the White Region Connectivity Case Study.

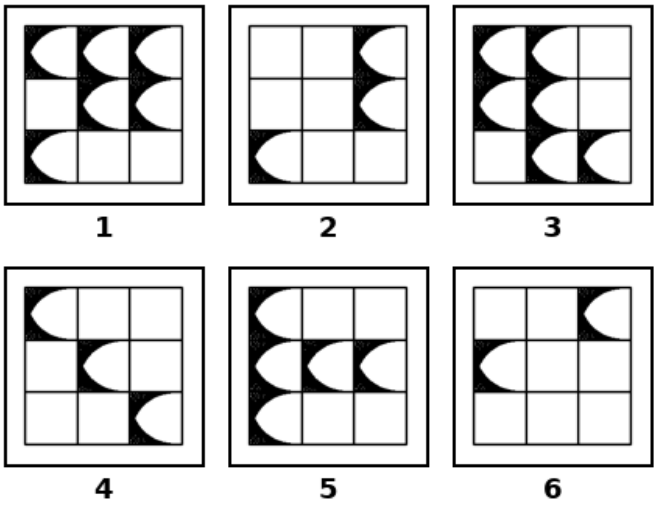
Case Study: Cluster 2 — Grid Occupancy Density

Question

Classify the following six images into two groups (three images per group) based on a common pattern, and indicate the numbers for each group.

Options:

A. {3, 4, 5}, {6, 2, 1} **B.** {1, 2, 4}, {5, 3, 6} **C.** {3, 2, 1}, {5, 4, 6} **D.** {6, 4, 2}, {5, 3, 1}



Logical Analysis

As illustrated in figure, each figure consists of a 3×3 grid with varying numbers of filled cells. The core pattern for classification is the occupancy density:

- **Low-Density Group (Figures 2, 4, 6):** These figures contain fewer filled cells, ranging from 2 to 3.
- **High-Density Group (Figures 1, 3, 5):** These figures contain a higher concentration of filled cells, ranging from 5 to 6.

By comparing the cell counts across all figures, the most consistent grouping aligns with the numerical thresholds of occupancy.

Correct Answer: D

Table 10: Overview of the Grid Occupancy Density Case Study.

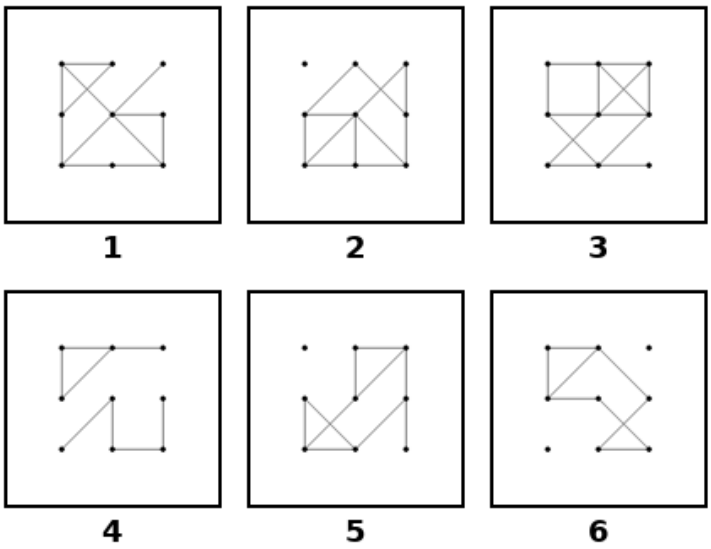
Case Study: Cluster 3 — Topological Connectivity

Question

Based on the common pattern among the figures, divide the six images into two groups of three, and identify the image numbers included in each group.

Options:

A. {1, 2, 4}, {3, 5, 6} **B.** {6, 2, 5}, {4, 1, 3} **C.** {1, 5, 4}, {3, 2, 6} **D.** {3, 4, 6}, {5, 1, 2}



Logical Analysis

As illustrated in figure, each figure consists of a 3×3 grid of points with connecting lines. The core pattern for grouping is the graph's connectivity:

- **Group I (Figures 1, 3, 4):** These figures are fully connected graphs where every point is part of a single interconnected component.
- **Group II (Figures 2, 5, 6):** These figures are not fully connected and contain isolated points or separate connected components.

Although the total number of lines varies (ranging from 8 to 14), connectivity remains the most consistent and fundamental grouping criterion.

Correct Answer: B

Table 11: Overview of the Topological Connectivity Case Study.

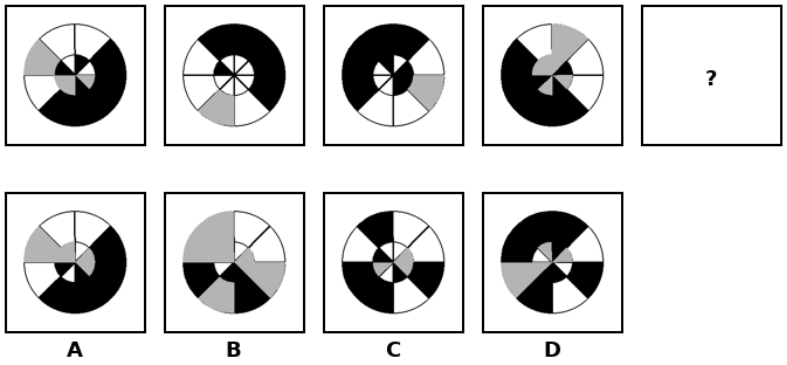
Case Study: Cluster 4 — Fixed-step Rotation Pattern

Question

Based on the pattern in the first four figures, select the fifth figure from the options that follows the established rule. Make your correct choice by identifying and applying the underlying pattern.

Options:

A. Option A **B.** Option B **C.** Option C **D.** Option D



Logical Analysis

As illustrated in figure, the sequence of the first four figures exhibits a regular transformation in the outer ring filling pattern:

- **Rotational Rule:** Starting from the first figure, the filled sector of the outer ring rotates **counterclockwise by 2 sectors** in each subsequent step.
- **Pattern Progression:** $Fig.1 \xrightarrow{+2\ sectors} Fig.2 \xrightarrow{+2\ sectors} Fig.3 \xrightarrow{+2\ sectors} Fig.4.$

To maintain this fixed-step rotation, the fifth figure must feature a filling pattern that is rotated another 2 sectors counterclockwise from Figure 4. This specific transformation aligns precisely with the configuration shown in Option A.

Correct Answer: A

Table 12: Overview of the Fixed-step Rotation Pattern Case Study.

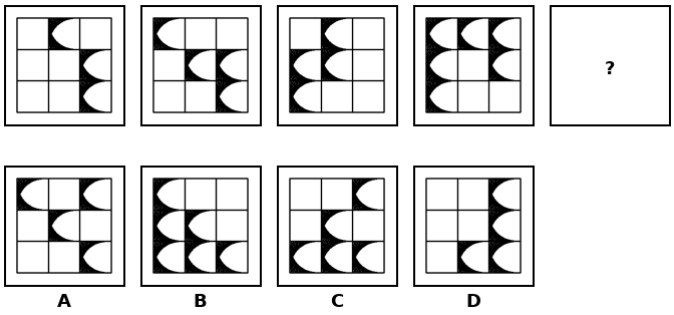
Case Study: Cluster 5 — Grid Column Progression

Question

Observe the pattern in the first four figures, and select the fifth figure from the following options that best fits this pattern.

Options:

A. Option A **B.** Option B **C.** Option C **D.** Option D



Logical Analysis

As illustrated in figure, the pattern in these 3×3 grids is determined by the specific number of filled positions in the leftmost column:

- **Quantitative Progression:** The number of filled positions in the first column increases strictly by 1 in each step ($0 \rightarrow 1 \rightarrow 2 \rightarrow 3$).
- **Inference:** Following this arithmetic progression, the first column in Figure 5 should maintain its fully occupied state (3 filled positions).

The model must identify that the leftmost column's density is the governing variable, leading to the selection of Option B.

Correct Answer: B

Table 13: Overview of the Grid Column Progression Case Study.

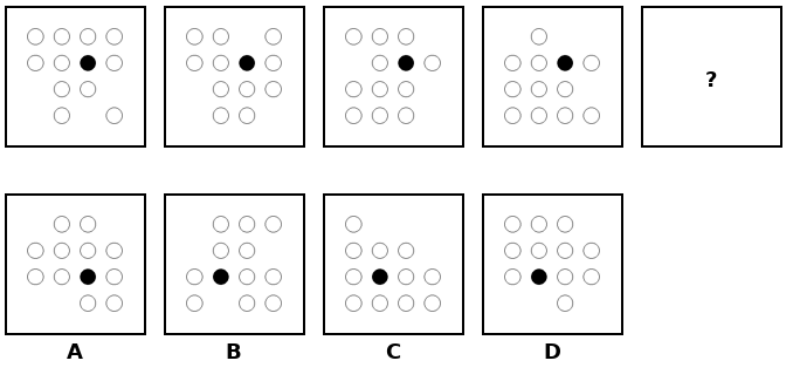
Case Study: Cluster 6 — Spatial Region Distribution Shift

Question

Observe the pattern presented in the first four figures, and select the fifth figure from the following options that follows this pattern.

Options:

A. Option A **B.** Option B **C.** Option C **D.** Option D



Logical Analysis

As illustrated in figure, the hidden circles follow a specific distributional shift between the upper half (0–7) and lower half (8–15):

- **Migration Pattern:** There is a linear migration from the lower region to the upper region: Fig. 1 (0:4) $\rightarrow$ Fig. 2 (1:3) $\rightarrow$ Fig. 3 (2:2) $\rightarrow$ Fig. 4 (3:1).
- **Trend Analysis:** The upper region count follows an increasing arithmetic sequence ($0 \rightarrow 1 \rightarrow 2 \rightarrow 3$).

Following this logic, Figure 5 must feature 4 hidden circles in the upper half and 0 in the lower half, which uniquely matches Option C.

Correct Answer: C

Table 14: Overview of the Spatial Region Distribution Shift Case Study.

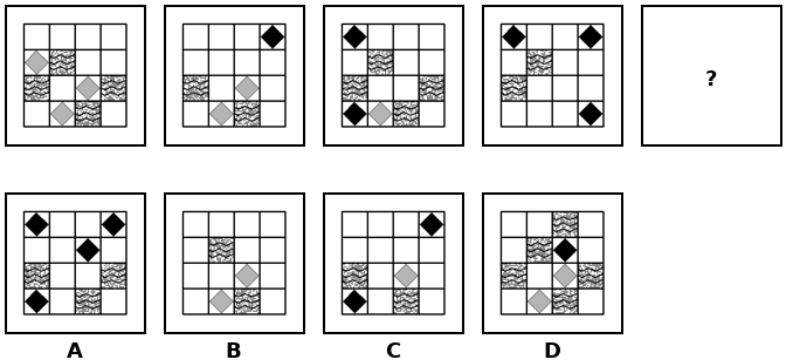
Case Study: Cluster 7 — Mirror-like Progression Pattern

Question

Observe the pattern in the first four figures and select the fifth figure from the following options that best fits the established rule. Make your correct choice based on the pattern you have identified.

Options:

A. Option A **B.** Option B **C.** Option C **D.** Option D



Logical Analysis

As illustrated in figure, the first four figures consist of 4×4 grids containing cells filled with black, gray, or wavy lines. The reasoning logic is derived from a complementary (mirror-like) numerical progression:

- **Black Cell Progression:** The count of black cells increases sequentially: $0 \rightarrow 1 \rightarrow 2 \rightarrow 3$.
- **Gray Cell Progression:** Conversely, the count of gray cells decreases sequentially: $3 \rightarrow 2 \rightarrow 1 \rightarrow 0$.

Following this established trend, the fifth figure is expected to contain exactly 4 black cells and 0 gray cells. This specific numerical relationship is uniquely satisfied by the configuration in Option A.

Correct Answer: A

Table 15: Overview of the Mirror-like Progression Case Study.

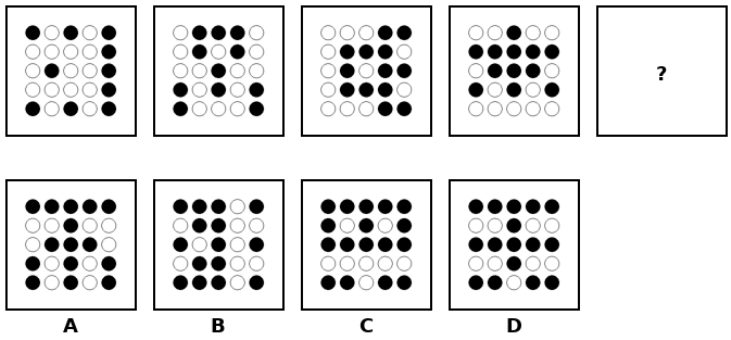
Case Study: Cluster 8 — Symmetry Alternation Pattern

Question

Observe the pattern in the first four figures, and select the fifth figure from the following options that best fits this pattern.

Options:

A. Option A **B.** Option B **C.** Option C **D.** Option D



Logical Analysis

As illustrated in figure, the reasoning logic is based on the alternating symmetry types of the sequential figures:

- **Sequence Observation:**
  - Figure 1: Horizontally symmetric.
  - Figure 2: Vertically symmetric.
  - Figure 3: Horizontally symmetric.
  - Figure 4: Vertically symmetric.
- **Rule Identification:** The symmetry type follows an alternating textA-B-A-B **pattern (Horizontal $\rightarrow$ Vertical)**.

According to this established periodic rule, the fifth figure in the sequence must return to the initial state, being textbfhorizontally symmetric. This requirement is uniquely satisfied by Option B.

Correct Answer: B

Table 16: Overview of the Symmetry Alternation Pattern Case Study.

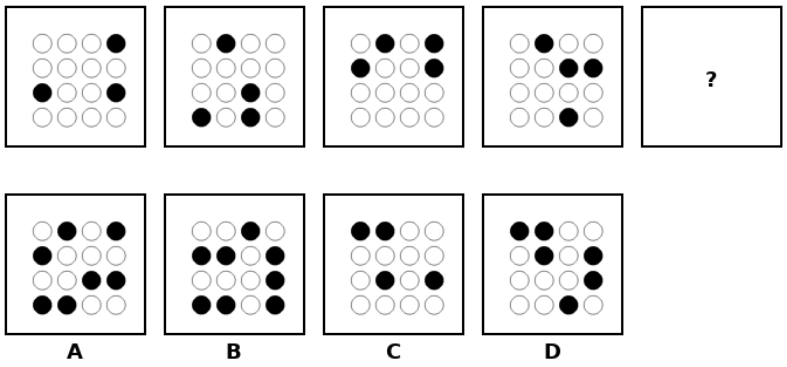
Case Study: Cluster 9 — Largest Connected Region Decrement Pattern

Question

Based on the pattern in the first four figures, select the fifth figure from the options that follows the same rule. Make the correct choice by identifying and applying the established pattern.

Options:

A. Option A **B.** Option B **C.** Option C **D.** Option D



Logical Analysis

As illustrated in figure, the reasoning pattern is derived from the numerical count of circles within the largest connected region of each grid:

- **Rule Identification:** The number of circles in the largest connected region follows a strictly decreasing arithmetic progression:
  - Figure 1: 13 circles.
  - Figure 2: 12 circles (decrement of 1).
  - Figure 3: 11 circles (decrement of 1).
  - Figure 4: 10 circles (decrement of 1).
- **Inference:** Following this decremental pattern ($13 \rightarrow 12 \rightarrow 11 \rightarrow 10$), the fifth figure must contain exactly **9 circles** within its largest connected region.

The placement of black circles progressively divides the grid into smaller components. This specific numerical requirement is only satisfied by Option D.

Correct Answer: D

Table 17: Overview of the Largest Connected Region Decrement Pattern Case Study.

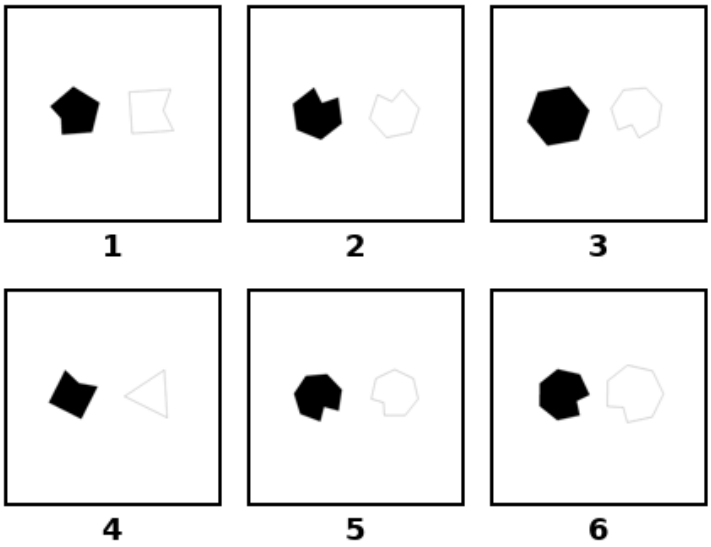
Case Study: Cluster 10 — Edge Parity and Quantitative Comparison

Question

Classify the following six images into two groups (three images per group) based on a common pattern, and indicate the numbers for each group.

Options:

A. {6, 5, 2}, {4, 1, 3} **B.** {6, 3, 1}, {4, 5, 2} **C.** {2, 6, 4}, {1, 5, 3} **D.** {6, 3, 2}, {4, 1, 5}



Logical Analysis

As illustrated in figure, the classification is determined by the numerical relationship between the black edges (N_{black}) and white edges (N_{white}) in each figure:

- **Group I (Figures 2, 5, 6):** These figures exhibit textbfEdge Equality, where the number of black edges equals the number of white edges ($N_{black} = N_{white}$).
  - Fig 2: 7 black, 7 white; Fig 5: 8 black, 8 white; Fig 6: 8 black, 8 white.
- **Group II (Figures 1, 3, 4):** These figures exhibit **Edge Disparity**, where the number of edges differs.
  - Fig 1: 6 black, 5 white (differ by 1); Fig 3: 6 black, 8 white; Fig 4: 5 black, 3 white.

Based on this quantitative pattern, Figures {6, 5, 2} and {4, 1, 3} form the two distinct categories, which aligns with Option A.

Correct Answer: A

Table 18: Overview of the Edge Parity and Quantitative Comparison Case Study.

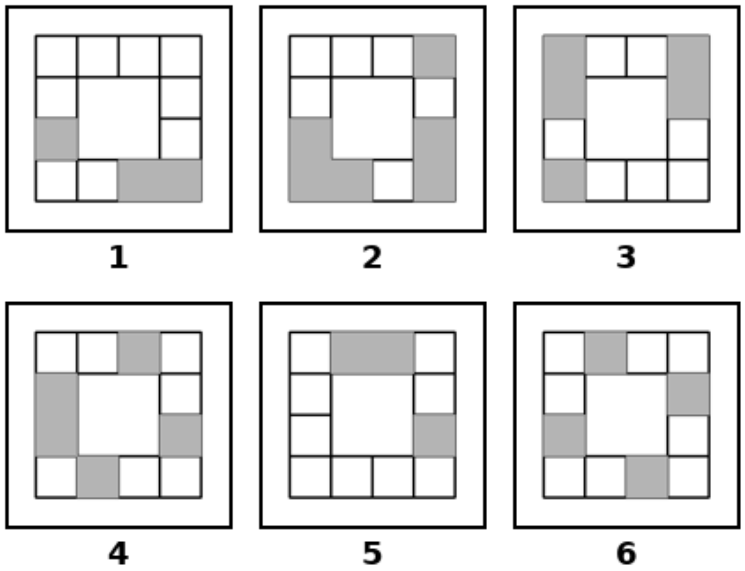
Case Study: Cluster 11 — Corner Occupancy Patterns

Question

Based on the common features of the figures, divide the six images into two groups, each containing three images, and identify the image numbers for each group. Choose the correct option according to the grouping pattern.

Options:

A. {4, 2, 6}, {3, 1, 5} **B.** {1, 6, 5}, {4, 3, 2} **C.** {6, 5, 4}, {2, 1, 3} **D.** {5, 2, 3}, {6, 1, 4}



Logical Analysis

As illustrated in figure, the six figures can be differentiated by observing the filling status of the four corner positions (top-left, top-right, bottom-left, and bottom-right) of the circular paths:

- **Group I (Figures 1, 2, 3):** These figures exhibit **Corner Occupancy**, where at least one of the four corner positions is filled with gray.
- **Group II (Figures 4, 5, 6):** These figures exhibit **Empty Corners**, where none of the four corner positions are filled.

By isolating the corner coordinates as the key diagnostic feature, Figures {6, 5, 4} are grouped together, and Figures {2, 1, 3} form the other group. This corresponds to the arrangement in Option C.

Correct Answer: C

Table 19: Overview of the Corner Occupancy Patterns Case Study.

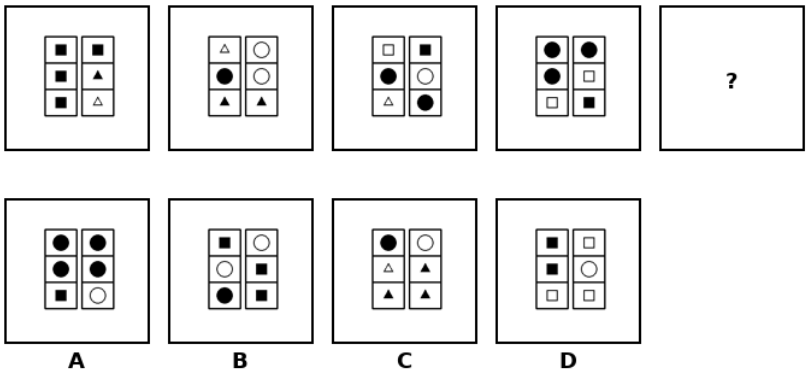
Case Study: Cluster 12 — Conveyor Belt Quantitative Pattern

Question

Observe the pattern in the first four figures and select the fifth figure from the options below that follows the established rule. Make your choice based on the pattern you have identified.

Options:

A. Option A **B.** Option B **C.** Option C **D.** Option D



Logical Analysis

As illustrated in figure, the grids represent dual conveyor belts containing mixed geometric shapes (circles, triangles, squares). The reasoning logic is determined by the cumulative count of black-filled circles across both columns:

- **Quantitative Rule:** The total number of black circles follows a linear arithmetic progression, increasing by 1 in each subsequent figure:
  - Fig 1: 0 black circles.
  - Fig 2: 1 black circle ($+1$).
  - Fig 3: 2 black circles ($+1$).
  - Fig 4: 3 black circles ($+1$).
- **Inference:** To maintain this constant growth rate ($0 \rightarrow 1 \rightarrow 2 \rightarrow 3$), the fifth figure must contain a total of exactly **4 black circles** across its left and right columns.

While other shapes (squares, triangles) vary in quantity, they do not follow a strictly linear progression, confirming that the black circle count is the governing attribute. Option A uniquely satisfies this numerical requirement.

Correct Answer: A

Table 20: Overview of the Conveyor Belt Quantitative Pattern Case Study.

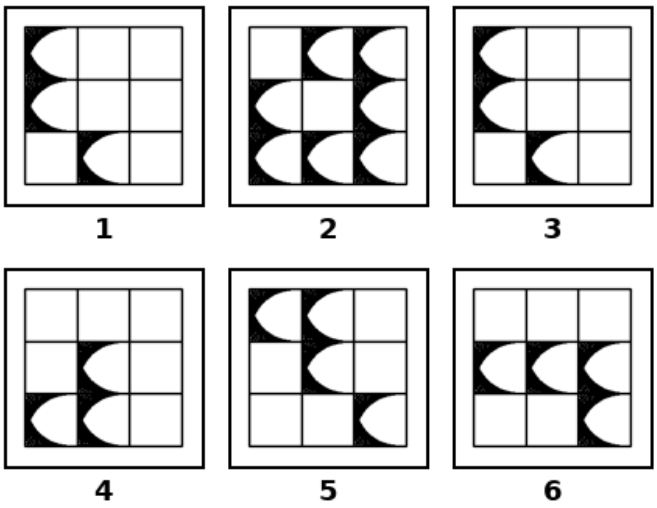
Case Study: Cluster 13 — Grid Centroid Occupancy

Question

Classify the following six images into two groups (three images per group) based on a common pattern, and indicate the numbers for each group.

Options:

A. {4, 2, 6}, {5, 3, 1} **B.** {3, 2, 1}, {5, 4, 6} **C.** {6, 2, 1}, {3, 4, 5} **D.** {6, 2, 5}, {1, 4, 3}



Logical Analysis

As illustrated in figure, each figure is a 3×3 grid. The primary feature for classification is the occupancy status of the center cell (centroid):

- **Group I (Figures 1, 2, 3):** In these figures, the center cell remains **unfilled**, regardless of the status of the surrounding eight cells.
- **Group II (Figures 4, 5, 6):** Conversely, these figures all feature a **filled** center cell, establishing a distinct topological category.

The classification logic abstracts away the complexity of peripheral cell patterns to focus on a singular, critical position. This binary distinction uniquely aligns with the groupings in Option B.

Correct Answer: B

Table 21: Overview of the Grid Centroid Occupancy Case Study.

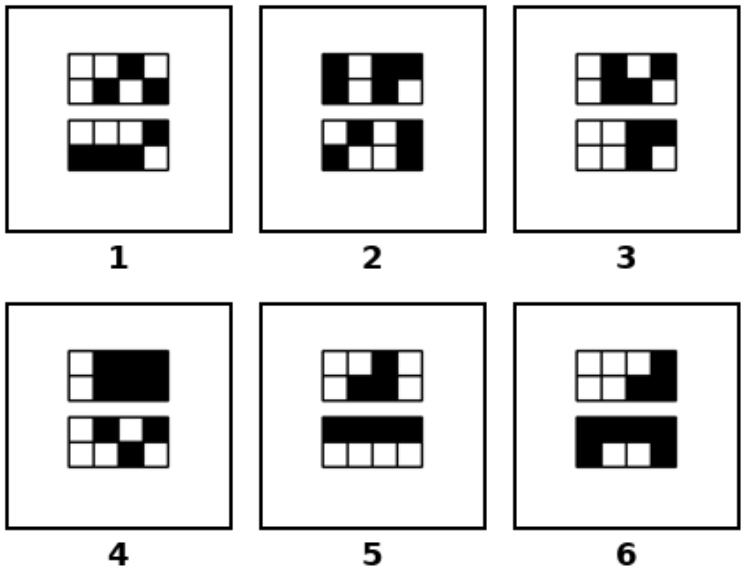
Case Study: Cluster 14 — Grid Area Quantitative Summations

Question

Divide the following six images into two groups based on a common pattern, with each group containing three images. Identify the corresponding numbers for each group.

Options:

A. {5, 1, 4}, {6, 3, 2} **B.** {1, 4, 2}, {3, 6, 5} **C.** {5, 2, 4}, {3, 6, 1} **D.** {2, 6, 4}, {3, 5, 1}



Logical Analysis

As illustrated in figure, each figure consists of a 2×4 grid structure divided into upper and lower sections. The classification variable is the total number of filled cells ($N_{total} = N_{upper} + N_{lower}$):

- **Group I (Figures 1, 3, 5):** These figures share a cumulative total of **7 filled cells**.
  - Fig 1: 3 (up) + 4 (low) = 7; Fig 3: 4 (up) + 3 (low) = 7; Fig 5: 3 (up) + 4 (low) = 7.
- **Group II (Figures 2, 4, 6):** These figures share a cumulative total of **9 filled cells**.
  - Fig 2: 5 (up) + 4 (low) = 9; Fig 4: 6 (up) + 3 (low) = 9; Fig 6: 3 (up) + 6 (low) = 9.

The reasoning logic remains consistent across the partitioning regardless of the specific distribution between the upper and lower rows. This grouping aligns with Option D.

Correct Answer: D

Table 22: Overview of the Grid Area Quantitative Case Study.