A Survey of Multimodal Mathematical Reasoning: From Perception, Alignment to Reasoning

Tianyu Yang[1*], **Sihong Wu**[2*], **Yilun Zhao**[2*], **Zhenwen Liang**[1], **Lisen Dai**[3], **Chen Zhao**[4], **Minhao Cheng**[5], **Arman Cohan**[2], **Xiangliang Zhang**[1†]

[1]University of Notre Dame [2]Yale University [3]Columbia University

[4]New York University [5]Pennsylvania State University

{tyang4, xzhang33}@nd.edu

Abstract

Multimodal Mathematical Reasoning (MMR) has recently attracted increasing attention for its capability to solve mathematical problems involving both textual and visual modalities. However, current models still face significant challenges in real-world visual math tasks, often misinterpreting diagrams, failing to align mathematical symbols with visual evidence, or producing inconsistent reasoning steps. Moreover, existing evaluations mainly focus on checking final answers rather than verifying the correctness or executability of each intermediate step. A growing body of recent research addresses these issues by integrating structured perception, explicit alignment, and verifiable reasoning within unified frameworks. To establish a clear roadmap for understanding and comparing different MMR approaches, we systematically review them around four fundamental questions: (1) What to extract from multimodal inputs, (2) How to represent and align textual and visual information, (3) How to perform the reasoning, and (4) How to evaluate the correctness of the overall reasoning process. Finally, we discuss open challenges and share our thoughts on future research directions. To keep pace with this rapidly evolving field, we maintain an open-source GitHub repository to continuously track ongoing progress in Multimodal Mathematical Reasoning.

1 Introduction

Large Language Models (LLMs) have recently advanced mathematical reasoning, achieving state-of-the-art results on various symbolic and arithmetic tasks, from elementary school level to college level (Guo et al., 2025a; DeepMind, 2024). However, in practice, mathematics often involves multimodal information. Many real-world problems in education (Ku et al., 2025), scientific discovery (Du et al., 2025), and interactive professional systems (Hu et al., 2024) require reasoning over visual structures and spatial relations. Solving these problems often requires interpreting diagrams, coordinate plots, charts, tables, and mixed-modality documents (Lu et al., 2021d; Saikh et al., 2022; Lee et al., 2023). In these contexts, visual elements encode critical constraints—such as incidence, parallelism, numeric scales, and layout semantics—that text-only models simply cannot perceive (Chen et al., 2025c).

To handle this complexity, a line of work focuses on integrating perception, symbolic understanding, and executable reasoning across modalities, defining the field of Multimodal Mathematical Reasoning (MMR) (Chen et al., 2021; Lu et al., 2021d; Saikh et al., 2022). Compared with purely text-based approaches (Lewkowycz et al., 2022; Liang et al., 2023), MMR approaches significantly improves evidence completeness by grounding visual cues. Nonetheless, these multimodal learning approaches substantially increase reasoning complexity: a model must jointly interpret visual cues, align them with symbolic expressions, and execute consistent multi-step reasoning across modalities (Sheng et al., 2025; Chen et al., 2021). This strong multimodal coupling introduces new, non-trivial challenges related to structured perception, cross-modal alignment, and verifiable reasoning. .

Given the importance of MMR and its rapid progress, we are motivated to present this survey that foregrounds fundamental mechanisms of addressing MMR using Multimodal LLMs (MLLMs). Prior efforts primarily catalog benchmarks and methodologies for MMR (Yan et al., 2024a) or discuss MLLM ecosystem roles (Reasoner, Enhancer, Planner) (Yan et al., 2024a). In contrast, we take a **vertical, process-centric view**: we articulate what is needed to solve MMR end-to-end and position MLLM-based approaches along this roadmap. Concretely, we organize the field around four questions: **1)** what to extract from multimodal inputs,

[*]Equal contributions.

[†]Correspondence

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 42386–42408

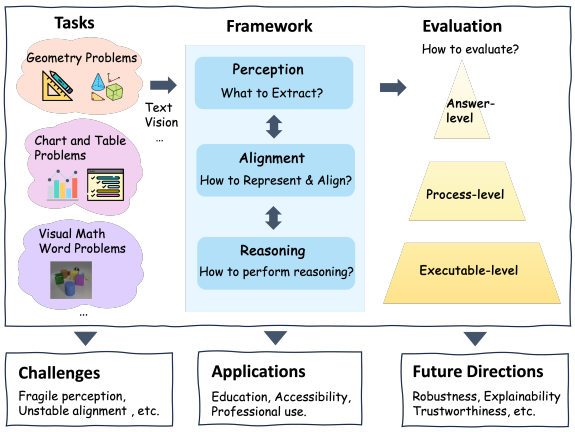

Figure 1: The roadmap of this survey.

2) how to represent and align textual and visual information, **3)** how to perform the reasoning (e.g., CoT, program-aided, tool use), and **4)** how to evaluate the correctness of the reasoning process. More discussion about our work vs related surveys is provided in Appendix A and Table 1.

Centered on these four questions, we organize MLLM-based MMR methods under a **Perception–Alignment–Reasoning** (PAR) framework, which decomposes MMR approaches into three interdependent stages: (1) **Perception**, extracting structured mathematical evidence from visual and textual modalities; (2) **Alignment**, mapping perceived facts to symbolic or executable representations; and (3) **Reasoning**, conducting interpretable and verifiable inference over the aligned representations (e.g., CoT, program execution, tool use).

To complement this process-centric perspective, we further introduce a companion evaluation hierarchy, the **Answer–Process–Executable** (APE) framework. APE assesses correctness at three levels, *answer* (task accuracy), *process* (faithfulness of intermediate reasoning steps), and *executable* (verification via executable checks). Together, PAR and APE provide a systematic lens for dissecting multimodal *mathematical* reasoning enabling both a comprehensive synthesis of prior work and a diagnostic understanding of where current MLLMs succeed or fail to reason faithfully.

The roadmap of our survey is shown in Figure 1. We begin by outlining the core challenges and preliminaries of MMR, including main task families and the structure of perception outputs. We then formalize the PAR pipeline and synthesize methods at each stage. For *Perception*, we track the path from symbolic parsers to pipelines built on large multimodal models (see Section 2). For *Alignment*, we cover executable intermediates, symbolic and neural hybrids, cross-modal alignment frameworks, and pretraining and finetuning strategies (see Section 3). For *Reasoning*, we review deliberate chains, reinforcement learning, tool-augmented and executable reasoning, and process feedback and verification (see Section 4). Next, we map major benchmarks and datasets to APE levels and to PAR stages (see Section 5), and we provide consolidated tables for direct comparison and diagnostic analysis (see Figure 2 or Tables 1-2). We finally conclude the survey by outlining open challenges and future directions (see Section 6).

2 Perception: What to Extract?

In the PAR framework (see Fig 2), perception addresses the first and central question, what to extract from multimodal inputs before alignment and reasoning can occur. Unlike generic vision tasks, mathematical perception must yield structured, computation relevant evidence rather than only objects or text. Given multimodal inputs, i.e., $X \subseteq \{T, D, C, I\}$ a mixture of text T, diagram D, chart or table C, and image I, the perception function $p : X \mapsto \mathcal{F}$ extracts a set of mathematical facts $\mathcal{F}$ spanning three levels: (i) low level primitives such as points, lines, axes, or objects, (ii) structural relations such as incidence, parallelism, axis series binding, or row and column layouts, and (iii) quantitative attributes such as lengths, angles, values, and units. Note that perception is essential; errors at this stage propagate downstream and can lead to misalignment or faulty reasoning.

To ground *PAR* in concrete settings, we introduce three representative task families: *geometry problems, chart/table problems*, and *visual math word problems*, as examples of what evidence needs to be extracted. We then summarize the task-oriented datasets through the lens of *PAR*, as detailed in Table 2, which provides the complete list of datasets for each task. Finally, we review the methodological evolution of perception, from symbolic parsers to neural encoders to LMM based pipelines, and conclude with an outlook on open challenges and promising directions.

Geometry Problems. Geometry problem solving requires models to jointly parse textual descriptions T and diagrams D to produce numerical values, symbolic relations, or complete proofs: $f : (T, D) \mapsto y$. Perception in this task focuses on recognizing geometric primitives such as points,

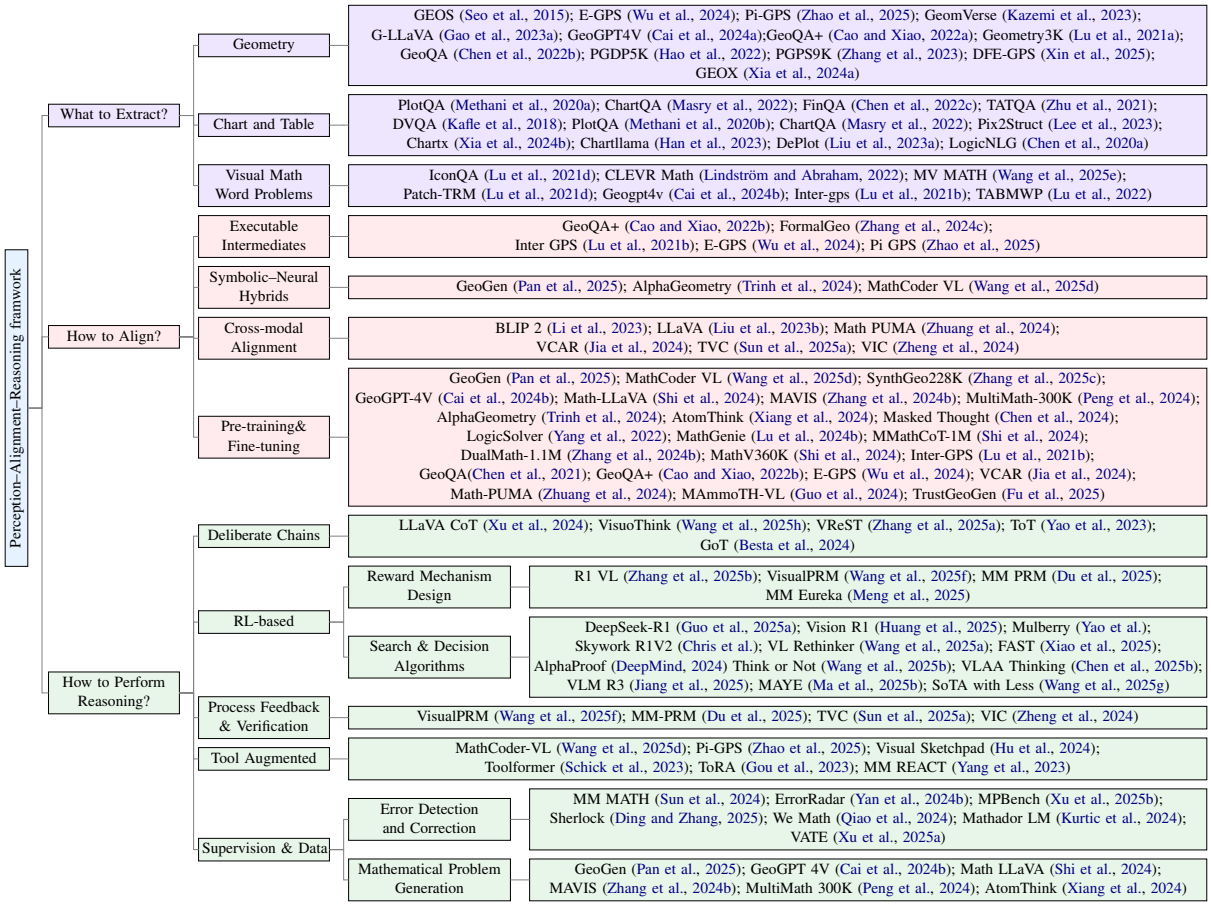

Figure 2: Taxonomy of Perception, Alignment and Reasoning framework.

lines, and angles, understanding their spatial relations, and grounding textual references to diagrammatic structures before performing deductive reasoning. Method development has progressed from symbolic theorem provers such as GEOS (Seo et al., 2015), to neural vision–language models, and more recently to hybrid pipelines with executable programs such as E-GPS (Wu et al., 2024) and Pi-GPS (Zhao et al., 2025), which enhance verifiability and explainability. LMMs further introduce a new perception paradigm, enabling both improved geometric understanding, as seen in GeomVerse (Kazemi et al., 2023), and large-scale synthetic data generation, as demonstrated by G-LLaVA (Gao et al., 2023a) and GeoGPT4V (Cai et al., 2024a). Recent work further explores *diagram formalization* and formal-language pretraining to improve structural understanding and robustness under domain shift, such as DFE-GPS (Xin et al., 2025) and GEOX (Xia et al., 2024a). Representative datasets include Geometry3K (Lu et al., 2021a), GeoQA and GeoQA+ (Chen et al., 2022b; Cao and Xiao, 2022a), PGDP5K (Hao et al., 2022), and PGPS9K (Zhang et al., 2023).

Chart and Table Problems. Chart and table problems assess the ability to interpret structured visual data C in response to a natural language query Q, formalized as $f : (C, Q) \mapsto a$, where a denotes the predicted answer. Models must accurately perceive visual layouts such as axes, legends, rows, and columns, ground linguistic references to these visual elements, and perform numerical or logical reasoning based on the extracted structure. Perception in this domain has evolved from explicit symbolic parsing (Kafle et al., 2018; Methani et al., 2020b; Masry et al., 2022) to neural vision–language models that jointly encode layout and text (Lee et al., 2023), and more recently to LMM-based instruction-tuned frameworks (Xia et al., 2024b; Han et al., 2023) that integrate structural perception with executable reasoning. DePlot (Liu et al., 2023a) and LogicNLG (Chen et al., 2020a) bridge perception and alignment through chart to table translation. Key benchmarks include PlotQA (Methani et al., 2020a), ChartQA (Masry et al., 2022), FinQA (Chen et al., 2022c), and TATQA (Zhu et al., 2021).

Visual Math Word Problems. Visual Math Word

Problems require solving natural-language math queries grounded in visual scenes: $f : (I, Q) \mapsto a$, where Q denotes the natural-language question and a denotes the predicted answer. Typical skills include object counting, attribute reasoning, quantity comparison, and cross-image co-reference. Methods have gradually shifted from symbolic perception and explicit object relation parsing like Patch-TRM (Lu et al., 2021d) to neural multimodal encoders that learn visual–textual correspondences (Lu et al., 2021b), and more recently to large multimodal models capable of holistic scene understanding and chain-of-thought reasoning (Cai et al., 2024b). Representative datasets include IconQA (Lu et al., 2021d), CLEVR-Math (Lindström and Abraham, 2022), TABMWP (Lu et al., 2022)and MV-MATH (Wang et al., 2025e).

Method Evolution and Outlook. Methods for mathematical perception have progressed from symbolic parsers and handcrafted rules to neural encoders that couple visual grounding with textual understanding, and now to LMMs unified through pretraining and instruction tuning. Despite their generality, LMMs often struggle with fine-grained perception, such as misreading geometric elements or chart layouts. Future work should focus on precise structure perception, executable supervision, and combining neural and symbolic reasoning for reliable results.

3 Alignment: How to Represent & Align?

Alignment bridges perception and reasoning. It defines how perceived visual facts are structured and mapped to symbolic or linguistic forms so that downstream reasoning becomes interpretable and verifiable. In mathematical contexts, alignment connects visual entities such as geometric primitives, chart axes, and table layouts with textual predicates or executable intermediates like geometry description languages, constraint sets, proof sketches, chart or table operators, SQL queries, and program-of-thought traces. The key challenge is to represent and align multimodal information while preserving symbolic fidelity and remaining robust to visual noise and domain variation. This section reviews alignment techniques from four complementary perspectives: (1) *executable intermediates* that formalize visual content into checkable programs, (2) *symbolic–neural hybrids* that couple neural perception with symbolic reasoning engines, (3) *cross-modal frameworks* that stabilize vision–language coupling, and (4) *pre-training and fine-tuning strategies* that provide large-scale priors and task-specific supervision.

3.1 Executable Intermediates

A key direction is converting visual content into formal, checkable intermediates that support symbolic reasoning. Inter-GPS (Lu et al., 2021b) annotate geometry problems with domain-specific languages to enable interpretable execution. E-GPS (Wu et al., 2024) integrates a symbolic solver with a diagram parser for verifiable step-by-step solutions. Pi-GPS (Zhao et al., 2025) introduces a multimodal rectifier to disambiguate diagrams before theorem-driven solving. R1-OneVision (Yang et al., 2025) scales this idea by transforming diagrams into textual formalizations for large-scale consistency training. Beyond geometry, chart and table reasoning convert visual marks into code- or SQL-like operators to ensure numeric correctness by design. Executable intermediates thus anchor alignment and make reasoning verifiable.

3.2 Symbolic–Neural Hybrids

Hybrid pipelines combine symbolic rigor with neural flexibility. GeoGen (Pan et al., 2025) aligns diagrams with executable programs under symbolic supervision. MathCoder-VL (Wang et al., 2025d) uses code-based cross-modal supervision to reinforce visual and text alignment and program-level faithfulness. AlphaGeometry (Trinh et al., 2024) integrates theorem libraries with neural search to handle complex geometric deductions. By injecting formal structure while retaining perceptual capacity, these hybrids enhance interpretability, transferability, and reasoning stability.

3.3 Cross-modal Alignment Frameworks

General frameworks provide reusable backbones for stable vision–language coupling. BLIP-2 (Li et al., 2023) links vision encoders to LLMs and serves as a base for math-specific extensions. LLaVA (Liu et al., 2023b) introduces instruction-following alignment for visual inputs. Math-PUMA (Zhuang et al., 2024) applies progressive staged alignment for long-chain stability, while VCAR (Jia et al., 2024) follows a "describe-then-reason" curriculum. For long-horizon reasoning, TVC (Sun et al., 2025a) maintains persistent visual conditioning, and VIC (Zheng et al., 2024) composes textual plans with late fusion to avoid drift. Curriculum- and conditioning-based designs help

reduce cumulative errors and stabilize multi-step reasoning.

3.4 Pre-training and Fine-tuning as Enablers

Large-scale pre-training provides broad coverage and alignment priors. Geo170K (Gao et al., 2023b), SynthGeo228K (Zhang et al., 2025c), TrustGeoGen (Fu et al., 2025) and GeoGPT-4V (Cai et al., 2024b) expand diagram–text coupling at scale. Math-LLaVA (Shi et al., 2024) and MAVIS (Zhang et al., 2024b) extend instruction-tuned data with visual reasoning. MultiMath-300K (Peng et al., 2024) contributes multimodal K–12 problems with stepwise annotations.Beyond these, MAmmoTH-VL (Guo et al., 2024) scales to 12M instruction pairs for multimodal pre-training, while (Fu et al., 2025) generates verified geometric data for reliable training. Symbolic resources like AlphaGeometry (Trinh et al., 2024) and auto-diagram construction (Krueger et al., 2021) further enhance formal priors. Objective design mixes grounding with process supervision—Masked Thought (Chen et al., 2024) learns from partial steps, LogicSolver (Yang et al., 2022) integrates logical constraints, and MathGenie (Lu et al., 2024b) generates synthetic CoT data.

Fine-tuning specializes alignment toward executable reasoning. MMathCoT-1M and DualMath-1.1M (Shi et al., 2024; Zhang et al., 2024b) link QA with dual-view trajectories, while MathV360K (Shi et al., 2024) and MAVIS (Zhang et al., 2024b) provide diagram-based instruction data. Datasets such as Geometry3K (Lu et al., 2021b), GeoQA (Chen et al., 2021), and E-GPS (Wu et al., 2024) enable symbolic supervision and program-level verifiability. Curricular designs like VCAR (Jia et al., 2024), Math-PUMA (Zhuang et al., 2024), and AtomThink (Xiang et al., 2024) progressively refine perception and reasoning, making alignment robust and transferable.

Outlook and Comparison. Executable intermediates ensure verifiability but are brittle under domain shifts. Symbolic–neural hybrids improve robustness yet add complexity. Cross-modal frameworks scale well but risk inconsistencies without explicit execution. Pre-training and fine-tuning bring generality but depend on data fidelity. In practice, combining executable precision, hybrid robustness, curriculum stability, and large-scale priors can perhaps achieve the best balance between reliability and generalization.

4 How to perform Reasoning?

After perception and alignment produce structured representations, the final stage concerns how models perform reliable inference. Reasoning in multimodal mathematical tasks involves executing stable and verifiable computation from structured inputs. Four paradigms dominate: (1) *Deliberate chain (e.g., CoT) methods*, which externalize intermediate steps to expose and guide reasoning; (2) *Reinforcement learning methods*, which optimize long-horizon decision sequences via reward-guided search; (3) *Tool-augmented reasoning*, which employs external solvers or code execution to enforce formal correctness; and (4) *Process feedback and verification*, which introduces critics or verifiers to assess intermediate steps (e.g., executable checks, self-consistency), improving validity and interpretability. These approaches collectively enhance robustness and faithfulness across long reasoning chains. Beyond these main paradigms, *Error Detection and Correction* (to flag and repair faulty traces) and *Mathematical Problem Generation* (to synthesize diverse, curriculum-aligned instances) play supportive roles that strengthen process supervision and dataset curation. Due to space limits, we defer discussion of these topics to Appendix C.

4.1 Deliberate Chains (e.g., CoT)

In-Context Learning (ICL) with multimodal chain-of-thought prompts models to externalize intermediate steps. LLaVA-CoT (Xu et al., 2024) shows that structured prompts can elicit more reliable reasoning paths. TVC (Sun et al., 2025a) injects persistent visual conditioning at every step to mitigate forgetting. VIC (Zheng et al., 2024) composes plans in text first and fuses vision later to reduce cross-modal drift. I2L (Wang et al., 2024b) embeds exemplars directly on the visual canvas to strengthen grounding. AtomThink (Xiang et al., 2024) decomposes reasoning into atomic steps, improving compositionality and enabling fine-grained supervision. These methods are lightweight and effective, but without strong grounding or verification mechanisms, they may still drift from evidence.

Beyond linear chains, Tree of Thoughts (ToT) (Yao et al., 2023) generalizes CoT by exploring and self-evaluating multiple branches of intermediate thoughts, and Graph of Thoughts (GoT) (Besta et al., 2024) further models non-linear dependencies among partial solutions. For multimodal set-

Benchmark	Year (Venue)	Eval Level	PAR Stage	Key Contributions
ChartQA (Masry et al., 2022)	2022 (ACL Findings)	Answer	Perception + Reasoning	Real charts; logical & numeric QA.
FigureQA (Kahou et al., 2017)	2018 (ICLR Workshop)	Answer	Perception	Synthetic charts; controlled reasoning.
PlotQA (Methani et al., 2020a)	2020 (WACV)	Answer	Perception + Reasoning	Real plots; open-vocab numeric answers.
IconQA (Lu et al., 2021d)	2021 (NeurIPS)	Answer	Perception + Reasoning	Large icon-based multimodal math.
CLEVR-Math (Lindström and Abraham, 2022)	2022 (NeSy Workshop)	Answer	Perception + Reasoning	Synthetic compositional arithmetic.
FinQA (Chen et al., 2022c)	2021 (EMNLP)	Answer	Alignment + Reasoning	Financial table-text; gold programs.
TAT-QA (Zhu et al., 2021)	2021 (ACL)	Answer	Alignment + Reasoning	Table-text numeracy in reports.
MultiHiertt (Zhao et al., 2022)	2022 (ACL)	Answer	Alignment + Reasoning	Financial table-text; gold programs.
DocMath-Eval (Zhao et al., 2024)	2024 (ACL)	Answer	Alignment + Reasoning	Financial table-text; gold evidence.
ChartQAPro (Masry et al., 2025)	2025 (ACL Findings)	Answer	Perception + Alignment	Harder charts incl. dashboards.
CharXiv (Wang et al., 2024d)	2024 (NeurIPS)	Answer	Perception	Human-curated arXiv charts.
MM-MATH (Sun et al., 2024)	2024 (EMNLP Findings)	Process	Reasoning	Step types & error labels.
MPBench (Xu et al., 2025b)	2025 (ACL Findings)	Process	Reasoning	PRM / step-judge benchmarking.
ErrorRadar (Yan et al., 2024b)	2024 (ICLR Workshop)	Process	Reasoning	Fine-grained error taxonomy.
Sherlock (Ding and Zhang, 2025)	2025 (NeurIPS)	Process	Reasoning	Multimodal error detect & repair.
We-Math (Qiao et al., 2024)	2025 (ACL)	Process	Reasoning	Principle-centered process probing.
MathVerse (Zhang et al., 2024a)	2024 (ECCV)	Process	All	Diagram perturbations; CoT step scoring.
CHAMP (Mao et al., 2024)	2024 (ACL Findings)	Process	Reasoning	Competition items; wrong-step tags.
PolyMATH (Gupta et al., 2024)	2024 (arXiv)	Process	Perception + Reasoning	Image–text puzzles; cognitive coverage.
GeoQA+ (Cao and Xiao, 2022b)	2022 (COLING)	Executable	Alignment + Reasoning	Geometry QA with executable programs.
Geometry3K (Lu et al., 2021a)	2021 (ACL)	Executable	Perception + Alignment	Dense formal language for geometry.
E-GPS (Lu et al., 2021b; Wu et al., 2024)	2024 (CVPR)	Executable	All	Solver+parser; verifiable steps.
FormalGeo (Zhang et al., 2024c)	2024 (MATH-AI)	Executable	Alignment + Reasoning	Olympiad-level formal proofs.
Pi-GPS (Zhao et al., 2025)	2025 (arXiv)	Executable	Alignment + Reasoning	Rectifier and solver for proofs.
WikiSQL (Zhong et al., 2017)	2017 (arxiv)	Executable	Alignment + Reasoning	NL→SQL with execution accuracy.
MathVista (Lu et al., 2024a)	2024 (ICLR)	Comprehensive	All	Aggregated multimodal suite.
MATH-V (Wang et al., 2024a)	2024 (NeurIPS)	Comprehensive	All	Difficulty-calibrated visual math.
OlympiadBench (He et al., 2024)	2024 (ACL)	Comprehensive	All	Bilingual competition-grade; stepwise.
MathScape (Liang et al., 2024a)	2024 (arXiv)	Comprehensive	All	Photo scenarios; multi-dim evaluation.
CMM-Math (Liu et al., 2024)	2024 (ACMMM)	Comprehensive	All	Chinese multimodal math.
Children's Olympiads (Cherian et al., 2024)	2024 (ESEM)	Comprehensive	All	Olympiad-style problems.
MM-PRM (Du et al., 2025)	2025 (arXiv)	Comprehensive	All	Real-world K-12 multimodal QA.

Table 1: Evaluation benchmarks organized by the APE hierarchy, aligned with corresponding PAR stages.

tings, AGoT (Yang et al., 2024) adapts GoT to multi-modal representation learning via an aggregation graph that soft-prompts and routes reasoning across aspects. For multimodal mathematical reasoning specifically, VisuoThink (Wang et al., 2025h) performs multimodal tree search with interleaved vision–text steps, and VReST (Zhang et al., 2025a) combines Monte Carlo Tree Search with a self-reward signal to deepen exploration and reports state-of-the-art results on several multimodal math benchmarks. Together, these ToT/GoT-style methods complement CoT by enabling branching, backtracking, and structured selection over intermediate solutions, which is particularly valuable for long-horizon visual–symbolic math problems.

4.2 RL-based Reasoning

Reinforcement learning (RL) approaches treat reasoning as a sequential decision process and optimize for long-horizon stability.

Reward Mechanism Design. R1-VL (Zhang et al., 2025b) introduces step-wise accuracy and validity rewards to encourage high-quality transitions. VisualPRM (Wang et al., 2025f) learns Process Reward Models (PRMs) from large-scale multimodal supervision to provide dense step-level feedback. MM-PRM (Du et al., 2025) combines PRM supervision with Monte Carlo Tree Search (MCTS) for comprehensive evaluation. MM-Eureka (Meng et al., 2025) explores rule-based RL to capture "visual aha" moments with minimal human annotation.

Search and Decision Algorithms. DeepSeek-R1 (Guo et al., 2025a) applies Group Relative Policy Optimization (GRPO) to jointly optimize reasoning and search, and Vision-R1 (Huang et al., 2025) extends this to multimodal settings. Mulberry (Yao et al.) integrates MCTS with reflective reasoning for iterative correction, while Skywork R1V2 (Chris et al.) combines Maximum a Posteriori Policy Optimization (MPO) and GRPO to balance detail and generalization. VL-Rethinker (Wang et al., 2025a) uses selective sample replay to mitigate vanishing advantages. FAST (Xiao et al., 2025) adapts inference depth to question complexity, and Think-or-Not? (Wang et al., 2025b) learns when to engage in deep reasoning. VLAA-Thinking (Chen et al., 2025b) studies reflection-aware optimization and contrasts RL with Supervised Fine-Tuning (SFT). VLM-R^3 (Jiang et al., 2025) proposes a three-stage pipeline of region recognition, reasoning, and refinement, while MAYE (Ma et al., 2025b) and SoTA-with-Less (Wang et al., 2025g) focus on sample efficiency via MCTS-guided data selection.

Beyond multimodal reasoning, AlphaProof (DeepMind, 2024) extends reinforcement learning to formal theorem proving via self-play and symbolic verification in Lean, achieving silver-medal performance on IMO problems. It exemplifies how RL can support verifiable and executable mathematical reasoning.

4.3 Tool-Augmented Reasoning

Tool-augmented methods delegate parts of reasoning to external symbolic systems or APIs to enhance modularity and correctness. Toolformer (Schick et al., 2023) demonstrates how LLMs can invoke external tools for symbolic computation and retrieval, while ToRA (Gou et al., 2023) organizes iterative loops of reasoning, tool calls, and result integration. COPRA (Thakur et al., 2023) composes multiple external capabilities adaptively, and MM-REACT (Yang et al., 2023) coordinates visual and textual tools for multimodal reasoning. For geometry, Visual Sketchpad (Hu et al., 2024) provides an interactive canvas that enables models to construct and reason visually, and Pi-GPS (Zhao et al., 2025) integrates parsers, verifiers, and symbolic solvers to produce provable results. Chameleon (Lu et al., 2023b) illustrates dynamic multi-tool composition, while MathCoder-VL (Wang et al., 2025d) uses code supervision to align diagrams with programs, making reasoning directly executable. Together, these systems show how tool integration supports structured, verifiable, and interpretable reasoning.

4.4 Process Feedback and Verification

VisualPRM (Wang et al., 2025f) provides process-level rewards that encourage valid steps and penalize errors. MM-PRM (Du et al., 2025) integrates PRM scoring with search, creating a generate–judge–revise loop for stable chains. Proof and program verifiers check intermediate Domain-Specific Language, code, or proof sketches, ensuring results are executable. At the representation level, TVC (Sun et al., 2025a) maintains visual conditioning during reasoning, while VIC (Zheng et al., 2024) reduces bias by text-first planning and late fusion. These approaches connect training with evaluation, ensuring that models are judged not only by answers but also by the correctness of their processes.

Outlook and Comparison. Different reasoning paradigms show complementary strengths. Deliberate chains are lightweight but risk drifting from visual evidence. Reinforcement learning stabilizes long reasoning yet demands costly rewards. Tool-augmented methods add modularity and verifiability but rely on stable interfaces. Process feedback improves auditability but needs dense supervision. Overall, hybrid systems combining explicit chains, selective RL, executable intermediates, and verification are most promising for robust, interpretable multimodal reasoning.

5 How to Evaluate?

To distinguish genuine mathematical reasoning from shortcut use, evaluation must span the full **PAR** pipeline and follow our **Answer–Process–Executable (APE)** hierarchy. **Answer**: Final-task metrics (e.g., accuracy) that are easy to report but can conflate perception errors (e.g., misread diagrams) and alignment errors (e.g., incorrect bindings) with reasoning mistakes. **Process**: Step-level checks that test whether intermediate reasoning is valid and *visually grounded* (i.e., consistent with extracted primitives and relations). **Executable**: Faithfulness via execution or proof checking (e.g., running code, verifying constraints/derivations) to directly assess alignment and reasoning correctness. We summarize how existing benchmarks map to the **APE** dimensions in Table 1. The table also covers **Comprehensive** benchmarks (see Appendix E) that combine diverse modalities, tasks, and difficulty levels to assess overall reasoning ability. Other benchmarks, including robustness (e.g., probing sensitivity to visual perturbations) and domain-specific sets (e.g., remote sensing), are discussed in Appendix D. Appendix G discusses comprehensive benchmarks, reviews robustness and domain-specific settings, and further discusses the interplay between datasets, models, and evaluation.

5.1 Answer-level Evaluation

Answer-level benchmarks judge the final answer with exact match or numeric tolerance. ChartQA (Masry et al., 2022) evaluates reasoning over diverse real-world charts; PlotQA (Methani et al., 2020a) stresses open-vocabulary and real-valued answers on scientific plots; FigureQA (Kahou et al., 2017) provides large-scale synthetic charts for controlled visual reasoning. IconQA (Lu et al., 2021d) assesses icon-like visual math with multiple formats and cognitive skills. CLEVR-Math (Lindström and Abraham, 2022) probes compositional

Dataset	Year (Venue)	PAR Stage	Size / Annotation	Key Contributions
Geometry Problem				
GEOS (Seo et al., 2015)	2015 (EMNLP)	Perception + Alignment	55 questions; text + diagram	early GPS baseline; text–diagram mapping
GEOS++ (Sachan et al., 2017)	2017 (EMNLP)	Alignment	1,406 questions; partial logical forms	SAT-style benchmark with logical grounding
Geometry3K (Lu et al., 2021c)	2021 (ACL)	Perception + Alignment	3,002 questions; dense formal language	formal grounding linking text and diagrams
GeoQA (Chen et al., 2022b)	2021 (ACL Findings)	Alignment + Reasoning	5,010 questions; executable programs	program-supervised QA
GeoQA+ (Cao and Xiao, 2022a)	2022 (COLING)	Alignment + Reasoning	extended set with harder steps	challenging multi-step reasoning test
PGDP5K (Hao et al., 2022)	2022 (IJCAI)	Perception	5,000 diagrams; primitive labels	dataset for geometric primitive parsing
PGPS9K (Zhang et al., 2023)	2023 (IJCAI)	All	9,022 items; fine-grained diagram + program	interpretable diagram–program pairs
UniGeo (Chen et al., 2022a)	2022 (EMNLP)	Alignment + Reasoning	4,998 calc + 9,543 proofs	unified format covering calculation and proof
GeomVerse (Kazemi et al., 2023)	2024 (ICML Workshop)	Reasoning	procedurally generated problems	synthetic benchmark to test reasoning capacity
FormalGeo7K (Zhang et al., 2024c)	2024 (NeurIPS Workshop)	Alignment + Reasoning	∼7,000 problems; diagram + formal solution	verifiable formal geometry tasks
Geo170K (Gao et al., 2023a)	2025 (ICLR)	Perception + Alignment	∼170,000 image–caption + QA pairs	large-scale geometry pretraining set
GeoGPT4V (Cai et al., 2024a)	2024 (EMNLP)	Perception + Alignment	4,900 synthesized + 19,000 mixed pairs	LLM-generated geometry text–figure dataset
MATHGLANCE (Sun et al., 2025c)	2025 (arXiv)	Perception	∼1,200 diagrams/1,600 questions; perception tags	isolates perception-level evaluation
Chart and Table Problems				
FigureQA (Kahou et al., 2018)	2018 (ICLR Workshop)	Perception	∼100,000 charts; ∼1M QA	synthetic chart reasoning dataset
DVQA (Kafle et al., 2018)	2018 (CVPR)	Perception	∼300,000 images; >3M QA	open vocabulary chart questions with metadata
PlotQA (Methani et al., 2020a)	2020 (WACV)	Perception	224,377 plots; ∼28.9M QA	real-valued numeric reasoning on scientific plots
ChartQA (Masry et al., 2022)	2022 (ACL Findings)	Perception + Alignment	9,600 human + 23,100 generated QA	visual + logical chart QA
CharXiv (Wang et al., 2024c)	2025 (NeurIPS)	Perception	2,323 curated charts	scientific chart understanding in real domain
ChartQAPro (Masry et al., 2025)	2025 (ACL)	Perception + Alignment	1,341 charts with dashboards	more complex visualization types
ChartQA-X (Hegde et al., 2025)	2025 (arxiv)	Alignment	30,299 charts with QA + rationale	supervision for explanation in charts
FinQA (Chen et al., 2022c)	2021 (EMNLP)	Alignment + Reasoning	8,281 cases with gold programs	hybrid table + text numerical reasoning
TAT-QA (Zhu et al., 2021)	2021 (ACL)	Alignment + Reasoning	16,552 QA in financial reports	table–text numerical reasoning benchmark
MultiHiertt (Zhao et al., 2022)	2022 (ACL)	Alignment + Reasoning	10,440 QAs in financial reports	hybrid table + text numerical reasoning
DocMath-Eval (Zhao et al., 2024)	2024 (ACL)	Alignment + Reasoning	4,000 QAs in financial reports; gold programs	hybrid table + text numerical reasoning
TabFact (Chen et al., 2020b)	2020 (ICLR)	Alignment	118,000 statements; 16,000 tables	table entailment verification dataset
WikiTableQuestions (Pasupat and Liang, 2015)	2015 (ACL)	Alignment + Reasoning	22,033 QA; 2,108 tables	compositional QA over web tables
WikiSQL (Zhong et al., 2017)	2017 (NeurIPS)	Alignment	80,654 NL–SQL; 24,241 tables	executable SQL supervision benchmark
DUDE (Landeghem et al., 2023)	2023 (ICCV)	All	multi-page document datasets	document-level reasoning with table/figure content
Visual Math Word Problems				
IconQA (Lu et al., 2021d)	2021 (NeurIPS)	Perception + Reasoning	107,439 questions; multiple formats	large-scale multimodal math QA benchmark
Icon645 (Lu et al., 2021d)	2021 (NeurIPS)	Perception	645,687 icons; 377 classes	icon pretraining resource
TABMWP (Lu et al., 2023c)	2023 (ICLR)	Alignment + Reasoning	38,431 problems; gold solutions / programs	table-based visual math word problems
CLEVR-Math (Lindström and Abraham, 2022)	2022 (NeSy)	Perception + Reasoning	synthetic image + text arithmetic	compositional arithmetic reasoning
MV-MATH (Wang et al., 2025e)	2025 (CVPR)	All	2,009 multi-image problems	cross-image dependency reasoning for K–12
MathVista (Lu et al., 2024a)	2024 (ICLR)	All	6,000+ visual math problems; 28 merged sets	combining diagrams, charts, and images
MATH-V (Wang et al., 2024a)	2024 (NeurIPS)	All	3,040 curated visual problems	higher-difficulty multimodal reasoning benchmark
Math2Visual (Wang et al., 2025c)	2024 (ACL Findings)	Perception + Alignment	12,000 generated visuals from math word text	benchmark for text-to-diagram generation in math

Table 2: Datasets grouped by task and annotated with the primary PAR stage they support, plus year, venue, size, and key contributions.

arithmetic in synthetic scenes. Hybrid table–text datasets such as FinQA (Chen et al., 2022c) and TAT-QA (Zhu et al., 2021) evaluate numerical reasoning over structured evidence. Answer-level evaluation is scalable and task-agnostic but cannot separate lucky guesses from correct reasoning, nor does it reveal where the Perception, Alignment and Reasoning pipeline failed.

5.2 Process-level Evaluation

Process-level benchmarks attach or elicit intermediate steps and score their validity, shifting the focus from answers to how solutions are produced. MM-MATH (Sun et al., 2024) provides step types and error annotations on middle-school problems with visual contexts. MPBench (Xu et al., 2025b) evaluates step-level judges and finds that many general multimodal models struggle with systematic error identification. ErrorRadar (Yan et al., 2024b) contributes fine-grained error taxonomies and labels for diagnostic analysis, and Sherlock (Ding and Zhang, 2025) extends multimodal process diagnosis with detailed failure categories. We-Math (Qiao et al., 2024) emphasizes principle-centered process evaluation beyond end-to-end scores, MathVerse (Zhang et al., 2024a) perturbs diagrams to test visual understanding beyond text priors, CHAMP (Mao et al., 2024) annotates concepts and hints and reports cases where models reach correct answers with wrong steps, and PolyMATH (Gupta et al., 2024) covers diverse cognitive categories including spatial and pattern reasoning. These resources enable audits of faithfulness and robustness while exposing where Perception or Alignment drifts translate into faulty Reasoning steps.

5.3 Executable-level Evaluation

Executable-level benchmarks require programs, proofs, or constraints that can be run or verified, directly testing symbolic Alignment and the faithfulness of Reasoning. GeoQA+ (Cao and Xiao, 2022b) annotates step-by-step programs for geometry and validates them by execution. FormalGeo (Zhang et al., 2024c) offers Olympiad-level geometry with formal statements, theorem sequences, and verifiable proofs. Inter-GPS and E-GPS (Lu et al., 2021b; Wu et al., 2024) provide formal languages and solver-backed pipelines, and Pi-GPS (Zhao et al., 2025) adds an LMM rectifier with a theorem-driven solver to produce provable chains. Executable metrics give clear pass or fail results that help identify alignment or reasoning errors, but they depend on reliable parsers and checkers.

6 Challenges and Future Directions

MMR has advanced rapidly, yet key challenges remain. Following the PAR framework, we summarize major limitations and future directions.

Perception. Current MLLMs show only a shallow understanding of visual information and often fail under layout or style changes (Liu et al., 2025a,b). Structured diagram parsing that captures primitives, topology, and layout improves robustness (Wu et al., 2024). A promising direction is to pair structured perception with formal interfaces such as code, proof sketches, or SQL, enabling visual evidence to be verified through execution (Zhao et al., 2025; Lu et al., 2021b).

Alignment. Fragmented *domain-specific languages (DSLs)* and inconsistent unit conventions cause misalignment and limit transfer. Future work should design unified, type-aware DSLs with explicit unit handling, constraint checking, and program verification (Pan et al., 2025) to standardize visual–symbolic mappings.

Reasoning. Long reasoning chains tend to drift from visual evidence. RL improves stability but is expensive and sensitive to reward design. Lightweight reward models, adaptive inference depth, and hybrid pipelines that delegate symbolic steps to external verifiers can reduce cost while maintaining robustness (Guo et al., 2025a; Huang et al., 2025; Wang et al., 2025a,f). This reflects a broader trade-off between stability and cost reinforcement learning enhances consistency but introduces heavy computational demands, motivating lightweight process rewards and symbolic verification for practical scalability.

However, benchmark-based evaluation remains limited: models may overfit to specific datasets or annotation styles rather than acquiring transferable reasoning skills. We provide a more systematic summary of practical failure patterns across perception, alignment, and reasoning in Appendix I. True reasoning should extend beyond curated benchmarks to unseen problems and open-ended contexts (Liang et al., 2024b; Cherian et al., 2024).

Future Opportunities. Applications such as intelligent tutoring, automated grading, and theorem explanation can enhance education through process-aware feedback (Ku et al., 2025; Du et al., 2025; Zhou et al., 2024). Accessibility tools like MathCAT and MathVision translate visual math into speech or braille with executable checks for accuracy (Soiffer, 2024; Awais et al., 2024). Professional systems for AR, VR, and engineering can integrate sketchpads, solvers, and code interfaces for verifiable design (Hu et al., 2024). Advancing these directions while addressing PAR-level challenges will lead to more reliable and interpretable multimodal reasoning systems. Detailed discussions on challenges and future opportunities are provided in Appendix F. We also provide practical design guidelines in Appendix H.

7 Conclusion

This paper presents a process-centered survey of MMR built on the Perception–Alignment–Reasoning (PAR) pipeline and the Answer–Process–Executable (APE) hierarchy. By organizing progress across geometry, chart and table reasoning, and visual math word problems, we show how structured perception, symbolic alignment, and verifiable reasoning jointly enable reliable multimodal intelligence. The PAR and APE frameworks offer a unified lens for understanding methods, benchmarks, and open issues, emphasizing structure-aware perception, executable intermediates, and process-level evaluation.

Acknowledgement

This work was supported by NSF Award No. 2333795.

Limitations

This survey primarily focuses on multimodal mathematical reasoning with visual–textual inputs. Although a few recent works explore broader modalities such as 3D reasoning, video understanding, or interactive tutoring, these studies remain relatively limited. We mention some of them for completeness but do not analyze them in depth. Given their growing importance, our future work may extend this survey or develop a dedicated review focusing on these emerging forms of dynamic and interactive multimodal reasoning to provide more comprehensive coverage for researchers.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Muhammad Awais, Tauqir Ahmed, Muhammad Aslam, Amjad Rehman, Faten S Alamri, Saeed Ali Bahaj, and Tanzila Saba. 2024. Mathvision: An accessible intelligent agent for visually impaired people to understand mathematical equations. *IEEE Access*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690.

Shihao Cai, Keqin Bao, Hangyu Guo, Jizhi Zhang, Jun Song, and Bo Zheng. 2024a. Geogpt4v: Towards geometric multi-modal large language models with geometric image generation. *Preprint*, arXiv:2406.11503.

Shihao Cai, Keqin Bao, Hangyu Guo, Jizhi Zhang, Jun Song, and Bo Zheng. 2024b. Geogpt4v: Towards geometric multi-modal large language models with geometric image generation. *arXiv preprint arXiv:2406.11503*.

Jie Cao and Jing Xiao. 2022a. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jie Cao and Jing Xiao. 2022b. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th international conference on computational linguistics*, pages 1511–1520.

Changyu Chen, Xiting Wang, Ting-En Lin, Ang Lv, Yuchuan Wu, Xin Gao, Ji-Rong Wen, Rui Yan, and Yongbin Li. 2024. Masked thought: Simply masking partial reasoning steps can improve mathematical reasoning learning of language models. *arXiv preprint arXiv:2403.02178*.

Felix Chen, Hangjie Yuan, Yunqiu Xu, Tao Feng, Jun Cen, Pengwei Liu, Zeying Huang, and Yi Yang. 2025a. Mathflow: Enhancing the perceptual flow of mllms for visual mathematical problems. *Preprint*, arXiv:2503.16549.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025b. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022a. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *Preprint*, arXiv:2212.02746.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. 2021. Geoqa: a geometric question answering benchmark towards multimodal numerical reasoning (2022). *URL https://arxiv. org/abs/2105.14517*, 40.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. 2022b. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *Preprint*, arXiv:2105.14517.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. 2025c. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *arXiv preprint arXiv:2503.01773*.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020a. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942, Online. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020b. Tabfact: A large-scale dataset for table-based fact verification. *Preprint*, arXiv:1909.02164.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2022c. Finqa: A dataset of numerical reasoning over financial data. *Preprint*, arXiv:2109.00122.

Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Joanna Matthiesen, Kevin Smith, and Josh Tenenbaum. 2024. Evaluating large vision-and-language models on children's mathematical olympiads. *Advances in Neural Information Processing Systems*, 37:15779–15800.

Y Wei Chris, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, Xuchen Song, and 1 others. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning, 2025. *URL https://arxiv.org/abs/2504.16656*.

DeepMind. 2024. Ai achieves silver-medal standard solving international mathematical olympiad problems. https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/. Accessed: 2025-10-06.

Yi Ding and Ruqi Zhang. 2025. Sherlock: Self-correcting reasoning in vision-language models. *arXiv preprint arXiv:2505.22651*.

Lingxiao Du, Fanqing Meng, Zongkai Liu, Zhixiang Zhou, Ping Luo, Qiaosheng Zhang, and Wenqi Shao.

2025. Mm-prm: Enhancing multimodal mathematical reasoning with scalable step-level supervision. *arXiv preprint arXiv:2505.13427*.

Daocheng Fu, Zijun Chen, Renqiu Xia, Qi Liu, Yuan Feng, Hongbin Zhou, Renrui Zhang, Shiyang Feng, Peng Gao, Junchi Yan, and 1 others. 2025. Trustgeogen: Scalable and formal-verified data engine for trustworthy multi-modal geometric problem solving. *arXiv preprint arXiv:2504.15780*.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. 2023a. G-llava: Solving geometric problem with multi-modal large language model. *Preprint*, arXiv:2312.11370.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and 1 others. 2023b. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. 2024. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*.

Zixian Guo, Ming Liu, Qilong Wang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2025b. Integrating visual interpretation and linguistic reasoning for math problem solving. *Preprint*, arXiv:2505.17609.

Himanshu Gupta, Shreyas Verma, Ujjwala Anantheswaran, Kevin Scaria, Mihir Parmar, Swaroop Mishra, and Chitta Baral. 2024. Polymath: A challenging multi-modal mathematical reasoning benchmark. *arXiv preprint arXiv:2410.14702*.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*.

Yihan Hao, Mingliang Zhang, Fei Yin, and Linlin Huang. 2022. Pgdp5k: A diagram parsing dataset for plane geometry problems. *Preprint*, arXiv:2205.09947.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Shamanthak Hegde, Pooyan Fazli, and Hasti Seifi. 2025. Chartqa-x: Generating explanations for visual chart reasoning. *Preprint*, arXiv:2504.13275.

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

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Mengzhao Jia, Zhihan Zhang, Wenhao Yu, Fangkai Jiao, and Meng Jiang. 2024. Describe-then-reason: Improving multimodal mathematical reasoning through visual comprehension training. *arXiv preprint arXiv:2404.14604*.

Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. 2025. Vlm-r3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *arXiv preprint arXiv:2505.16192*.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. 2018. Figureqa: An annotated figure dataset for visual reasoning. *Preprint*, arXiv:1710.07300.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. 2023. Geomverse: A systematic evaluation of large models for geometric reasoning. *Preprint*, arXiv:2312.12241.

Ryan Krueger, Jesse Michael Han, and Daniel Selsam. 2021. Automatically building diagrams for olympiad geometry problems. In *CADE*, pages 577–588.

Max Ku, Thomas Chong, Jonathan Leung, Krish Shah, Alvin Yu, and Wenhu Chen. 2025. Theoremexplainagent: Towards video-based multimodal explanations for llm theorem understanding. *arXiv preprint arXiv:2502.19400*.

Eldar Kurtic, Amir Moeini, and Dan Alistarh. 2024. Mathador-lm: A dynamic benchmark for mathematical reasoning on large language models. *arXiv preprint arXiv:2406.12572*.

Jordy Van Landeghem, Rubén Tito, Łukasz Borchmann, Michał Pietruszka, Paweł Józiak, Rafał Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Ackaert, Ernest Valveny, Matthew Blaschko, Sien Moens, and Tomasz Stanisławek. 2023. Document understanding dataset and evaluation (dude). *Preprint*, arXiv:2305.08455.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.

Chenglin Li, Qianglong Chen, Zhi Li, Feng Tao, and Yin Zhang. 2024. Vcbench: A controllable benchmark for symbolic and abstract challenges in video cognition. *arXiv preprint arXiv:2411.09105*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, and 1 others. 2025. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*.

Hao Liang, Linzhuang Sun, Minxuan Zhou, Zirong Chen, Meiyi Qiang, Mingan Lin, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. 2024a. Mathscape: Benchmarking multimodal large language models in real-world mathematical contexts. *arXiv e-prints*, pages arXiv–2408.

Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. 2024b. Scemqa: A scientific college entrance level multimodal question answering benchmark. *arXiv preprint arXiv:2402.05138*.

Zhenwen Liang, Tianyu Yang, Jipeng Zhang, and Xiangliang Zhang. 2023. Unimath: A foundational and multimodal mathematical reasoner. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7126–7133.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *Preprint*, arXiv:2208.05358.

Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023a. Deplot: One-shot visual language reasoning by plot-to-table translation. *Preprint*, arXiv:2212.10505.

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

Wentao Liu, Qianjun Pan, Yi Zhang, Zhuo Liu, Ji Wu, Jie Zhou, Aimin Zhou, Qin Chen, Bo Jiang, and Liang He. 2024. Cmm-math: A chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models. *arXiv preprint arXiv:2409.02834*.

Yufang Liu, Yao Du, Tao Ji, Jianing Wang, Yang Liu, Yuanbin Wu, Aimin Zhou, Mengdi Zhang, and Xunliang Cai. 2025a. The role of visual modality in multimodal mathematical reasoning: Challenges and insights. *Preprint*, arXiv:2503.04167.

Yufang Liu, Yao Du, Tao Ji, Jianing Wang, Yang Liu, Yuanbin Wu, Aimin Zhou, Mengdi Zhang, and Xunliang Cai. 2025b. The role of visual modality in multimodal mathematical reasoning: Challenges and insights. *arXiv preprint arXiv:2503.04167*.

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024a. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *Preprint*, arXiv:2310.02255.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021a. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786, Online. Association for Computational Linguistics.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021b. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021c. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *Preprint*, arXiv:2105.04165.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023b. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36:43447–43478.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023c. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *Preprint*, arXiv:2209.14610.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021d. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023d. A survey of deep learning for mathematical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14605–14631, Toronto, Canada. Association for Computational Linguistics.

Zimu Lu, Aojun Zhou, Houxing Ren, Ke Wang, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. 2024b. Mathgenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of llms. *arXiv preprint arXiv:2402.16352*.

Hongxu Ma, Chenbo Zhang, Lu Zhang, Jiaogen Zhou, Jihong Guan, and Shuigeng Zhou. 2025a. Fine-grained zero-shot object detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4504–4513.

Yan Ma, Steffi Chern, Xuyang Shen, Yiran Zhong, and Pengfei Liu. 2025b. Rethinking rl scaling for vision language models: A transparent, from-scratch framework and comprehensive evaluation scheme. *arXiv preprint arXiv:2504.02587*.

Yujun Mao, Yoon Kim, and Yilun Zhou. 2024. Champ: A competition-level dataset for fine-grained analyses of llms' mathematical reasoning capabilities. *arXiv preprint arXiv:2401.06961*.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2025. ChartQAPro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19123–19151, Vienna, Austria. Association for Computational Linguistics.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, and 1 others. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020a. Plotqa: Reasoning over scientific plots. *Preprint*, arXiv:1909.00997.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020b. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536.

Yicheng Pan, Zhenrong Zhang, Pengfei Hu, Jiefeng Ma, Jun Du, Jianshu Zhang, Quan Liu, Jianqing Gao, and Feng Ma. 2025. Enhancing the geometric problem-solving ability of multimodal llms via symbolic-neural integration. *arXiv preprint arXiv:2504.12773*.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. *Preprint*, arXiv:1508.00305.

Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. 2024. Multimath: Bridging visual and mathematical reasoning for large language models. *arXiv preprint arXiv:2409.00147*.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, and 1 others. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Mrinmaya Sachan, Kumar Dubey, and Eric Xing. 2017. From textbooks to knowledge: A case study in harvesting axiomatic knowledge from textbooks to solve geometry problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 773–784, Copenhagen, Denmark. Association for Computational Linguistics.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1466–1476, Lisbon, Portugal. Association for Computational Linguistics.

Jiayi Sheng, Luna Lyu, Jikai Jin, Tony Xia, Alex Gu, James Zou, and Pan Lu. 2025. Solving inequality proofs with large language models. *arXiv preprint arXiv:2506.07927*.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*.

Neil Soiffer. 2024. Mathcat: Math capable assistive technology.

Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. 2025a. Mitigating visual forgetting via take-along visual conditioning for multi-modal long cot reasoning. *arXiv preprint arXiv:2503.13360*.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Yuan Wu, Qi Liu, and 15 others. 2025b. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Comput. Surv.*, 57(11).

Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. 2024. Mm-math: Advancing multimodal math evaluation with process evaluation and fine-grained classification. *arXiv preprint arXiv:2404.05091*.

Yanpeng Sun, Shan Zhang, Wei Tang, Aotian Chen, Piotr Koniusz, Kai Zou, Yuan Xue, and Anton van den Hengel. 2025c. Mathglance: Multimodal large language models do not know where to look in mathematical diagrams. *Preprint*, arXiv:2503.20745.

Amitayush Thakur, George Tsoukalas, Yeming Wen, Jimmy Xin, and Swarat Chaudhuri. 2023. An in-context learning agent for formal theorem-proving. *arXiv preprint arXiv:2310.04353*.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025a. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*.

Jiaqi Wang, Kevin Qinghong Lin, James Cheng, and Mike Zheng Shou. 2025b. Think or not? selective reasoning via reinforcement learning for vision-language models. *arXiv preprint arXiv:2505.16854*.

Junling Wang, Anna Rutkiewicz, April Wang, and Mrinmaya Sachan. 2025c. Generating pedagogically meaningful visuals for math word problems: A new benchmark and analysis of text-to-image models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11229–11257, Vienna, Austria. Association for Computational Linguistics.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Ke Wang, Junting Pan, Linda Wei, Aojun Zhou, Weikang Shi, Zimu Lu, Han Xiao, Yunqiao Yang, Houxing Ren, Mingjie Zhan, and 1 others. 2025d. Mathcoder-vl: Bridging vision and code for enhanced multimodal mathematical reasoning. *arXiv preprint arXiv:2505.10557*.

Lei Wang, Wanyu Xu, Zhiqiang Hu, Yihuai Lan, Shan Dong, Hao Wang, Roy Ka-Wei Lee, and Ee-Peng Lim. 2024b. All in an aggregated image for in-image learning. *arXiv preprint arXiv:2402.17971*.

Peijie Wang, Zhong-Zhi Li, Fei Yin, Dekang Ran, and Cheng-Lin Liu. 2025e. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19541–19551.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, and 1 others. 2025f. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025g. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*.

Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. 2025h. Visuothink: Empowering lvlm reasoning with multimodal tree search. *arXiv preprint arXiv:2504.09130*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024c. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Preprint*, arXiv:2406.18521.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024d. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Wenjun Wu, Lingling Zhang, Jun Liu, Xi Tang, Yaxian Wang, Shaowei Wang, and Qianying Wang. 2024. E-gps: Explainable geometry problem solving via top-down solver and bottom-up generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13828–13837.

Renqiu Xia, Mingsheng Li, Hancheng Ye, Wenjie Wu, Hongbin Zhou, Jiakang Yuan, Tianshuo Peng, Xinyu Cai, Xiangchao Yan, Bin Wang, and 1 others. 2024a. Geox: Geometric problem solving through unified formalized vision-language pre-training. *arXiv preprint arXiv:2412.11863*.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Peng Ye, Min Dou, Botian Shi, and 1 others. 2024b. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*.

Kun Xiang, Zhili Liu, Zihao Jiang, Yunshuang Nie, Runhui Huang, Haoxiang Fan, Hanhui Li, Weiran Huang, Yihan Zeng, Jianhua Han, and 1 others. 2024. Atomthink: A slow thinking framework for multimodal mathematical reasoning. *arXiv preprint arXiv:2411.11930*.

Wenyi Xiao, Leilei Gan, Weilong Dai, Wanggui He, Ziwei Huang, Haoyuan Li, Fangxun Shu, Zhelun Yu, Peng Zhang, Hao Jiang, and 1 others. 2025. Fast-slow thinking for large vision-language model reasoning. *arXiv preprint arXiv:2504.18458*.

Yue Xin, Wenyuan Wang, Rui Pan, Ruida Wang, Howard Meng, Renjie Pi, Shizhe Diao, and Tong Zhang. 2025. Generalizable geometric image caption synthesis. *arXiv preprint arXiv:2509.15217*.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Shijia Xu, Yu Wang, Xiaolong Jia, Zhou Wu, Kai Liu, and April Xiaowen Dong. 2026a. Rcbsf: A multi-agent framework for automated contract revision via stackelberg game. *Preprint*, arXiv:2604.10740.

Shijia Xu, Zhou Wu, Xiaolong Jia, Yu Wang, Kai Liu, and April Xiaowen Dong. 2026b. Self-correcting rag: Enhancing faithfulness via mmkp context selection and nli-guided mcts. *Preprint*, arXiv:2604.10734.

Tianlong Xu, YiFan Zhang, Zhendong Chu, Shen Wang, and Qingsong Wen. 2025a. Ai-driven virtual teacher for enhanced educational efficiency: Leveraging large pretrain models for autonomous error analysis and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28801–28809.

Zhaopan Xu, Pengfei Zhou, Jiaxin Ai, Wangbo Zhao, Kai Wang, Xiaojiang Peng, Wenqi Shao, Hongxun Yao, and Kaipeng Zhang. 2025b. *arXiv preprint arXiv:2503.12505*.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2024a. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. *arXiv preprint arXiv:2412.11936*.

Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, and 1 others. 2024b. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. *arXiv preprint arXiv:2410.04509*.

Juncheng Yang, Zuchao Li, Shuai Xie, Wei Yu, Shijun Li, and Bo Du. 2024. Soft-prompting with graph-of-thought for multi-modal representation learning. *arXiv preprint arXiv:2404.04538*.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, and 1 others. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Zhicheng Yang, Jinghui Qin, Jiaqi Chen, Liang Lin, and Xiaodan Liang. 2022. Logicsolver: Towards interpretable math word problem solving with logical prompt-enhanced learning. *arXiv preprint arXiv:2205.08232*.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and 1 others. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search, 2024. *URL https://arxiv. org/abs/2412.18319*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Congzhi Zhang, Jiawei Peng, Zhenglin Wang, Yilong Lai, Haowen Sun, Heng Chang, Fei Ma, and

Weijiang Yu. 2025a. Vrest: Enhancing reasoning in large vision-language models through tree search and self-reward mechanism. *arXiv preprint arXiv:2506.08691*.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025b. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*.

Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. 2023. A multi-modal neural geometric solver with textual clauses parsed from diagram. *Preprint*, arXiv:2302.11097.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024a. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, and 1 others. 2024b. Mavis: Mathematical visual instruction tuning with an automatic data engine. *arXiv preprint arXiv:2407.08739*.

Xiaokai Zhang, Na Zhu, Yiming He, Jia Zou, Qike Huang, Xiaoxiao Jin, Yanjun Guo, Chenyang Mao, Yang Li, Zhe Zhu, Dengfeng Yue, Fangzhen Zhu, Yifan Wang, Yiwen Huang, Runan Wang, Cheng Qin, Zhenbing Zeng, Shaorong Xie, Xiangfeng Luo, and Tuo Leng. 2024c. Formalgeo: An extensible formalized framework for olympiad geometric problem solving. *Preprint*, arXiv:2310.18021.

Zeren Zhang, Jo-Ku Cheng, Jingyang Deng, Lu Tian, Jinwen Ma, Ziran Qin, Xiaokai Zhang, Na Zhu, and Tuo Leng. 2025c. Diagram formalization enhanced multi-modal geometry problem solver. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Junbo Zhao, Ting Zhang, Jiayu Sun, Mi Tian, and Hua Huang. 2025. Pi-gps: Enhancing geometry problem solving by unleashing the power of diagrammatic information. *arXiv preprint arXiv:2503.05543*.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600, Dublin, Ireland. Association for Computational Linguistics.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2024. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16103–16120, Bangkok, Thailand. Association for Computational Linguistics.

Haojie Zheng, Tianyang Xu, Hanchi Sun, Shu Pu, Ruoxi Chen, and Lichao Sun. 2024. Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination. *arXiv preprint arXiv:2411.12591*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *Preprint*, arXiv:1709.00103.

Minxuan Zhou, Hao Liang, Tianpeng Li, Zhiyu Wu, Mingan Lin, Linzhuang Sun, Yaqi Zhou, Yan Zhang, Xiaoqin Huang, Yicong Chen, and 1 others. 2024. Mathscape: Evaluating mllms in multimodal math scenarios through a hierarchical benchmark. *arXiv preprint arXiv:2408.07543*.

Yue Zhou, Litong Feng, Mengcheng Lan, Yiping Ke, Xue Jiang, and Wayne Zhang. Geomath: A benchmark for multimodal mathematical reasoning in remote sensing.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *Preprint*, arXiv:2105.07624.

Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin Zeng. 2024. Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. *Preprint*, arXiv:2408.08640.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2024. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*.

Appendix

A Related Surveys

As shown in Table 1, we summary recent related surveys. Recent surveys have examined mathematical reasoning and multimodal intelligence from complementary perspectives but differ in focus and depth. Lu et al. (2023d) reviewed deep learning for mathematical reasoning, summarizing architectures and datasets in the pre-LLM era but without multimodal or process-level analysis. Sun et al. (2025b) broadly discussed reasoning with foundation models across commonsense, logical, and mathematical domains, yet its treatment of symbolic and multimodal reasoning remains superficial. Ahn et al. (2024) analyzed LLM-based mathematical reasoning through four dimensions: tasks, methods, factors, and challenges, offering a structured text-centered view but overlooking visual grounding and reasoning processes. Yan et al. (2024a) extended this to the multimodal large language model (MLLM) era, organizing research by benchmarks, methodologies, and challenges, and introducing model roles as Reasoner, Enhancer, and Planner. However, its emphasis lies on ecosystem taxonomy rather than the internal mechanism connecting perception and symbolic alignment. Li et al. (2025) surveyed large multimodal reasoning models (LMRMs) and proposed a developmental roadmap from modular perception to agentic reasoning, integrating reinforcement learning and multimodal chain-of-thought. Although comprehensive in scope, it treats mathematics as one application and lacks formal analysis of symbolic-numeric grounding or verifiability.

In contrast, our survey focuses specifically on multimodal mathematical reasoning (MMR), abstracting the workflow into the Perception–Alignment–Reasoning (PAR) framework and the Answer–Process–Executable (APE) evaluation hierarchy. Together, PAR and APE provide a unified lens for understanding how multimodal evidence is perceived, aligned, and executed in verifiable reasoning. This framework bridges the symbolic–neural perspective of early deep learning, the text-based view of LLM reasoning, and the model-centric paradigm of MLLMs, offering the first process-level synthesis of multimodal mathematical reasoning.

Overall, previous surveys remain largely descriptive and domain-specific, while ours advances toward a process-level, verifiable, and multimodal understanding of mathematical reasoning that integrates perception, alignment, and reasoning within a coherent analytical framework.

B Reasoning Pipeline: Perception, Alignment and Reasoning

We abstract multimodal math reasoning into three stages. This view clarifies where systems fail and how to design robust solutions.

Perception. The goal is to recover computationally relevant visual facts. In geometry this means primitives and topology such as points, lines, angles, incidence, and equality. In charts and tables this means axes, legends, marks, tick reading, cell structure, and semantic units. Robust OCR and layout also matter in document settings. Errors at this stage, such as missed intersections or misread scales, often cascade.

Alignment. The next step is to bind visual facts to textual predicates or to an intermediate representation that can be executed. Examples include a geometry description language, a set of constraints, a proof language, a sequence of operators for charts and tables, a SQL query, or a program of thought trace. Alignment benefits from explicit anchors and structural losses, from code or program supervision, and from formal interfaces. To reduce cross modal drift during long chains of thought, recent strategies first compose reasoning in text and then consult visual evidence, or maintain visual conditioning throughout the chain.

Reasoning. The final step executes arithmetic, logic, theorem sequences, or programs, often with tool use such as calculators, symbolic solvers, or retrieval. Process level critics and rewards and search methods such as best of N or tree search help maintain validity over long chains. Retaining visual evidence and controlling bias are important for stability. In geometry, staged planning with verifier backed steps is especially effective.

This decomposition also guides evaluation. Some benchmarks focus on perception and alignment such as chart reading or primitive extraction. Others emphasize executable and checkable inference such as geometric proofs or program execution.

C Supervision and Data for Reasoning

C.1 Error Detection and Correction

In multimodal mathematical reasoning, inference often involves long chains of cross-modal steps, which requires not only evaluating the final answer

Survey	Venue & Year	Scope / Focus	Models	Focus
Lu et al. (2023d)	ACL'23	DL4Math	Deep Learning	Pre-LLM; model architectures and datasets;
Sun et al. (2025b)	ACM Computing'25	FM4Reason	MLLM	Broad reasoning (limited math/symbolic depth)
Ahn et al. (2024)	EACL Workshop'24	LLM4Math	LLM	Text-centered (non-MMR)
Yan et al. (2024a)	ACL Findings'25	MLLM4Math	MLLM	Benchmark- and Model-centric taxonomy
Li et al. (2025)	arXiv'25	LMRM	LLM/MLLM	Roadmap- and Stage-centric analysis
Ours	-	**MLLM4Math**	**MLLM**	**First unified process-level framework revealing internal mechanisms of multimodal mathematical reasoning**

Table 1: Comparisons between representative surveys and ours. "Models" column indicates model scope discussed in each survey (e.g., deep learning models, LLM, MLLM).

but also supervising and revising intermediate reasoning states. VisualPRM (Wang et al., 2025f) provides process-level rewards with dense supervision, encouraging valid reasoning transitions and penalizing deviations. MM-PRM (Du et al., 2025) integrates PRM scoring with Monte Carlo Tree Search to form a generate–judge–revise loop that stabilizes long reasoning chains. Mathador-LM (Kurtic et al., 2024) instantiates critique-driven revision for math solutions, promoting self-correction during inference. VATE (Xu et al., 2025a) targets classroom drafts with interactive feedback loops aligned with human pedagogy. Sherlock (Ding and Zhang, 2025) contributes fine-grained error taxonomies for process diagnosis, and ErrorRadar (Yan et al., 2024b; Ma et al., 2025a) provides labeled categories to localize typical failure modes. MM-MATH (Sun et al., 2024) supplies large-scale step and error annotations, while MPBench (Xu et al., 2025b) shows that general-purpose multimodal models still struggle with systematic error identification. Together, these systems and resources operationalize step-level judging and correction, so models are evaluated and improved by how they reason, not just by final answers.

C.2 Mathematical Problem Generation

In multimodal mathematical reasoning, generating high-quality problems is essential for driving model training and evaluation, especially by supplying process- and execution-level testbeds for perception, alignment, and reasoning. GeoGen (Pan et al., 2025) follows a generate–solve–verify loop coupling symbolic solvers with natural-language verbalization to guarantee checkable solutions. GeoGPT-4V (Cai et al., 2024b) co-generates aligned text–figure pairs with a strong multimodal model to broaden geometric coverage. Math-LLaVA with MathV360K (Shi et al., 2024) extends instruction-style data toward visual math, and MAVIS (Zhang et al., 2024b) provides an automatic data engine with chain-of-thought supervision for large-scale synthesis. MultiMath-300K (Peng et al., 2024) curates K–12 multimodal problems with captions and stepwise solutions for process-aware training. AtomThink (Xiang et al., 2024) offers long atomic chains of thought to supervise compositional reasoning, while MathCoder-VL (Wang et al., 2025d) uses code as supervision to align diagrams with executable programs for verifiable generation. These generation pipelines and corpora supply controllable, diverse, and executable data that strengthen perception and alignment while furnishing robust evaluation environments.

D Robustness and Domain-specific Benchmarks

Robustness benchmarks probe sensitivity to visual perturbations, multi-image dependencies, and domain shifts beyond standard evaluation. VCBench (Li et al., 2024) focuses on explicit multi-image reasoning dependencies. DynaMath (Zou et al., 2024) applies dynamic perturbations to test shortcut reliance. HC-M3D (Liu et al., 2025b) constructs near-duplicate images that flip correct answers to measure vision dependence. SMART-840 (Cherian et al., 2024) collects K–12 visuo-linguistic problems to assess fundamental multimodal skills under varied conditions. Domain specific sets such as GeoMath (Zhou et al.) target remote-sensing imagery and subject-specific math tasks, while MV-MATH (Wang et al., 2025e) extends multi-image reasoning to K–12 contexts. Together these datasets assess model stability, generalization, and cross-domain transfer for multimodal mathematical reasoning.

E Comprehensive Benchmarks

Comprehensive suites mix modalities, tasks, and difficulties to profile broad capabilities. MathVista (Lu et al., 2024a) aggregates problems from many sources spanning natural images, diagrams,

and charts. MATH-V (Wang et al., 2024a) emphasizes difficulty calibration and curated coverage across subjects. SceMQA (Liang et al., 2024b) introduces a scientific multimodal QA benchmark at the college entrance level including Mathematics and other core subjects to evaluate reasoning across disciplines. MM-K12 (Du et al., 2025) targets K–12 education scenarios with verifiable multimodal problems, bridging visual understanding and curriculum-level reasoning. OlympiadBench (Cherian et al., 2024) reports expert-level annotations enabling stepwise evaluation on competition-grade math and physics, while the Children's Olympiads benchmark (He et al., 2024) evaluates reasoning on competition problems designed for younger students. MathScape (Liang et al., 2024a) focuses on photo-based scenarios with hierarchical categories and multidimensional evaluation. CMM-Math (Liu et al., 2024) extends these benchmarks to the Chinese language setting, highlighting multilingual reasoning capabilities. These suites provide breadth and coverage but often entangle perception, alignment, and reasoning in a single score.

F Challenges and Future Directions

F.1 Challenges

Evaluation Challenges. While the proposed Answer–Process–Executable (APE) evaluation level provides a structured lens for assessing reasoning fidelity, the executable-level evaluation remains challenging to scale. Current executable benchmarks such as GeoQA+ (Cao and Xiao, 2022b), FormalGeo (Zhang et al., 2024c), and Pi-GPS (Zhao et al., 2025) depend on domain-specific languages, symbolic solvers, or theorem checkers that are largely confined to geometry or table reasoning tasks. Generalizing these pipelines to broader multimodal reasoning such as chart interpretation, visual word problems, or scientific document understanding requires unified annotation protocols and lightweight verification schemes. Moreover, executable evaluation often introduces heavy computational costs and relies on manually curated programs or proofs, limiting its practicality for large-scale MLLM assessment. Future work may explore scalable formal interfaces and semi-automated checkers that balance verifiability, coverage, and efficiency within the APE framework.

Cross-cutting Challenges. Data contamination, limited reproducibility, safety, and interpretability remain persistent issues. Leakage audits, standardized reporting, and verifier-backed pipelines can improve reliability. Executable intermediates, process judges, and proof or code verification support interpretability and trustworthy reasoning (Hu et al., 2024).

F.2 Future Opportunities

Multimodal mathematical reasoning enables diverse downstream applications that benefit from the model's ability to process and integrate visual and symbolic modalities. We categorize representative applications into three core areas:

1. Education and Learning. Education applications benefit greatly from multimodal reasoning. For example, in STEM learning, tools like TheoremExplainAgent (Ku et al., 2025) visually and symbolically guide students through theorems and problem-solving processes. Intelligent tutoring systems (Du et al., 2025) dynamically adapt based on student input, providing feedback by analyzing both diagrams and text. Automated grading systems (Zhou et al., 2024) can assess multi-step, visual-rich student solutions, improving evaluation accuracy and scalability.

2. Accessibility and Inclusivity. For learners with disabilities, multimodal reasoning systems enable accessible content delivery. MathCAT (Soiffer, 2024) and Mathvision (Awais et al., 2024) translate visual math into speech and braille, facilitating interaction with geometry or charts. These systems also support alternative input/output modalities (e.g., voice, haptics), ensuring inclusive engagement with mathematical content.

3. Professional and Interactive Systems. In real-world problem-solving tasks—such as data analysis, architecture, or engineering—professionals must reason over both visual schematics and textual instructions. Multimodal reasoning aids this integration. In parallel, interactive interfaces in AR/VR environments (Hu et al., 2024) allow users to engage with math through gestures, voice commands, or immersive visual aids. These interfaces, when empowered by multimodal reasoning, enhance spatial understanding and application-specific interaction.

G Interplay between Datasets, Models, and Evaluation.

The PAR and APE frameworks imply that datasets, model architectures, and evaluation pro-

tocols are not independent choices. What a benchmark annotates, and at which APE level, largely determines which stage of the Perception–Alignment–Reasoning pipeline is stressed; in turn, emerging modeling paradigms reveal gaps in existing benchmarks. Answer-only suites such as MathVista (Lu et al., 2023a) and MATH-V (Wang et al., 2024a) mainly report final accuracy on static diagrams, charts, and scenes. Under this setting, models often combine one-shot perception with generic CoT or program-of-thought decoding, and answer-level RL can already improve aggregate scores, but perception, alignment, and reasoning failures are entangled and shortcut strategies remain hard to diagnose.

Process-oriented and robustness benchmarks make these interactions more explicit. We-Math (Qiao et al., 2024) decomposes problems into concept-level sub-questions and reports IK/IG/CM/RM metrics, directly probing where knowledge and generalization fail along the reasoning chain. MathVerse (Zhang et al., 2024a) and related variants perturb diagrams or isolate text-only views to test whether models truly rely on visual evidence rather than textual priors. FlowVerse (Chen et al., 2025a) further factorizes problem information into DI/EI/RP/OQ versions and introduces FlowVerse-CoT-E, tying evaluation to step-level reasoning grounded in perceptual information. Dynamic benchmarks such as DynaMath (Zou et al., 2024) complement this by generating multiple visual and textual variants per seed question and comparing average- vs worst-case accuracy, emphasizing robustness under benign perturbations rather than single-shot success. Together with process-annotated corpora such as MultiMath-300K (Peng et al., 2024), these resources naturally favor step-aware supervision (e.g., PRMs, RL with process or outcome rewards, search-based refinement) and make Perception, Alignment, and Reasoning errors more observable. Executable- or program-level supervision further pushes models toward modular pipelines. Geometry datasets with DSLs, proofs, and solver-backed checks support systems that first convert diagrams into executable representations before reasoning. MathFlow and FlowVerse exemplify this trend in visual math: FlowVerse exposes which parts of a solution depend on perception versus abstract reasoning, and MathFlow decouples a dedicated perception module from a flexible inference LLM, showing that strengthening PAR's Perception stage can improve performance across many backbones. Decoupled frameworks such as DVLR (Guo et al., 2025b) similarly separate visual interpretation from linguistic reasoning and adopt outcome-rewarded joint tuning on geometry benchmarks, while RL methods like VL-Rethinker (Wang et al., 2025a) illustrate how, once process- and robustness-oriented benchmarks exist, self-reflective and perception-aware training strategies become natural responses. Viewed through PAR and APE, future benchmark design and model design should be co-planned: answer-only suites are still useful for breadth, but sustained progress will depend on more process-rich, dynamic, and executable benchmarks that expose failure modes at each PAR stage and support verifiable, visually grounded reasoning.

H Practical Design Guidelines

While our survey is organized along the PAR (Perception–Alignment–Reasoning) pipeline and the APE (Answer–Process–Executable) hierarchy, practitioners ultimately need concrete guidance on how to instantiate these abstractions in real systems. This subsection distills several practical design guidelines from the methods and benchmarks reviewed above and summarizes them as actionable take-home messages.

No universal optimal design. A central observation of this survey is that there is no "one-size-fits-all" multimodal mathematical reasoner. Executable, symbol-heavy pipelines provide strong guarantees and debuggability but are fragile to noisy perception and expensive to annotate. In contrast, purely neural, latent pipelines offer flexibility and robustness to imperfect inputs, yet make it difficult to enforce or inspect the underlying mathematical structure. Similarly, always-on deep reasoning (e.g., search, RL, and intensive tool augmentation) can improve robustness on difficult instances, but may be unnecessary or even harmful for routine, low-stakes problems due to increased latency and potential overfitting to benchmark-specific reward signals.

Choosing APE Levels and Benchmarks. For large-scale, low-stakes applications such as homework assistance or interactive practice, answer-level evaluation on broad suites like MathVista (Lu et al., 2023a) or ChartQA (Masry et al., 2022) is often sufficient to guide model selection, provided

that occasional errors are acceptable and qualitative inspection is used to detect obvious shortcut behavior. In safety-critical or high-stakes settings (e.g., automatic grading, high-level examinations, or formal theorem proving), process- or executable-level benchmarks—such as We-Math (Qiao et al., 2024), MM-MATH (Sun et al., 2024), FlowVerse (Chen et al., 2025a), NL2SQL-style datasets (Zhong et al., 2017), or formal geometry corpora—are preferable because they reveal where the reasoning chain fails and allow automatic verification of intermediate states. A practical rule of thumb is: (1) rely primarily on answer-level evaluation when coverage, scale, and latency are the dominant constraints and individual mistakes are tolerable; (2) adopt process-level evaluation when diagnosing typical failure modes (knowledge gaps, hallucinated steps, perception mistakes) is important; (3) favor executable-level evaluation when correctness and debuggability outweigh annotation cost and domain coverage.

Guidelines for Alignment Design. When verifiability and fine-grained error analysis are paramount—for instance, in exam grading or systems that must provide legally or pedagogically reliable feedback—executable or DSL-based alignment (e.g., geometry DSLs, SQL, program-of-thought operators) combined with solver-backed checks is preferable, despite higher engineering and annotation overhead. For broad, latency-sensitive platforms such as large-scale tutoring systems, lightweight latent alignment with unified abstractions on top of generic MLLM backbones is often more appropriate, trading strict guarantees for robustness to noisy diagrams and lower maintenance cost. Hybrid designs that use executable alignment for a small set of core skills (e.g., Euclidean geometry, table/SQL reasoning) and latent alignment elsewhere provide a pragmatic compromise when both formal guarantees and wide task coverage are required.

Guidelines for Reasoning Paradigms. For routine, low-stakes tasks, CoT-only or single-pass reasoning is typically adequate: such approaches are easy to deploy, respect strict latency budgets, and can be combined with simple calibration to reduce over-confident failures. For competition-level, research-style, or grading-style problems, RL-enhanced or search-based reasoning, often coupled with tool augmentation (e.g., calculators, theorem provers, program execution), is more suitable, as it prioritizes robustness and faithfulness over runtime. When both efficiency and reliability matter, selective or budgeted "think-more-when-needed" strategies form a practical middle ground: the model uses fast CoT for most inputs but automatically triggers deeper search or external tools on uncertain or adversarial cases, as indicated by uncertainty measures or self-consistency checks.

Recommended Configurations. Putting these pieces together, several patterns emerge as practically useful design recipes: (1) Safety-critical grading and assessment: executable or DSL-based alignment with solver-backed checks, combined with search-based or RL-enhanced reasoning, evaluated predominantly at process or executable APE levels. (2) Large-scale tutoring and practice platforms: latent alignment with unified representations and fast CoT-style or shallow multi-step reasoning, primarily evaluated at the answer level, with spot checks on process-level benchmarks. (3) Interactive tools balancing guarantees and responsiveness: hybrid alignment (symbolic for a core subset of tasks, latent elsewhere) together with selective or budgeted multi-step or tool-augmented reasoning, evaluated with a mix of answer-, process-, and executable-level benchmarks.

I Systematic Failure Patterns in Practical Settings

While Table 1 maps existing benchmarks to the Answer–Process–Executable (APE) hierarchy and the PAR stages, practical reliability also depends on how models fail in realistic conditions. Beyond aggregate scores, process-level, robustness, and comprehensive benchmarks expose recurring failure patterns that cut across perception, alignment, and reasoning. In this subsection, we synthesize these patterns along the PAR and APE dimensions, with a particular focus on sensitivity to low-quality diagrams, ambiguous multimodal references, and domain shifts between educational and scientific contexts.

Perception-level Failures. Models exhibit sensitivity to low-quality diagrams, including low resolution, compression artifacts, cluttered layouts, partial crops, and imperfect OCR such as handwritten annotations. The manifestations are task-dependent: in geometry, small perturbations lead to missed intersections, distorted angles, or misdetected primitives; in chart and table reasoning,

they surface as axis, legend, and scale extraction errors; in visual math word problems, they obscure small objects or local relations. Robustness-oriented resources such as VCBench (Li et al., 2024), DynaMath (Zou et al., 2024), HC-M3D (Liu et al., 2025a), and SMART-840 (Cherian et al., 2024) explicitly probe these sensitivities through multi-image dependencies, visual perturbations, and near-duplicate cases, while domain-specific sets like GeoMath and multi-image K–12 MV-MATH (Zhou et al.; Wang et al., 2025e) add further perception stressors in scientific and educational contexts.

Alignment-level Failures. A second class arises from ambiguous multimodal references and domain shifts between educational and scientific contexts. Errors include binding textual mentions such as "this triangle," "the bar for 2021," or "region A" to wrong regions, and unit or scale mismatches in charts and tables even when local perception is correct. Benchmarks such as ChartQA (Masry et al., 2022), PlotQA (Methani et al., 2020b), FinQA (Chen et al., 2022c), TAT-QA (Zhu et al., 2021), ChartQAPro (Masry et al., 2025), and CharXiv (Wang et al., 2024c) consistently reveal mistakes in axis and legend binding and unit normalization. Distributional differences between MM-K12 (Du et al., 2025), OlympiadBench (He et al., 2024), and Children's Olympiads (Cherian et al., 2024) versus scientific or photo-based suites such as MathScape (Zhou et al., 2024) and SceMQA (Liang et al., 2024b) further cause executable descriptions that appear valid to encode wrong bindings or mismatched assumptions, reducing transfer across settings.

Reasoning-level Failures. Even with mostly correct perception and alignment, models often produce unfaithful or brittle chains (Xu et al., 2026a). Process-level evaluations show cases where models reach correct answers via unsupported steps, hallucinated operations not grounded in visuals, or sharp drops on out-of-distribution problems despite plausible narratives (Xu et al., 2026b). Datasets such as MM-MATH (Sun et al., 2024), MPBench (Xu et al., 2025b), ErrorRadar (Yan et al., 2024b), Sherlock (Ding and Zhang, 2025), We-Math (Qiao et al., 2024), MathVerse (Zhang et al., 2024a), CHAMP (Mao et al., 2024), and Poly-MATH (Gupta et al., 2024) expose over-reliance on language priors, under-use of visual evidence, and gaps between answer-level success and process-level faithfulness. Executable resources including GeoQA+ (Cao and Xiao, 2022b), Geometry3K (Lu et al., 2021c), E-GPS (Wu et al., 2024), and Formal-Geo (Zhang et al., 2024c) further reveal reasoning traces that fail strict program or proof checking despite coherent text, highlighting latent misalignments and logical inconsistencies.

Findings. Viewed through PAR and APE, these patterns indicate that reliable deployment requires perception robust to degraded or stylistically varied diagrams, alignment that handles ambiguous references and cross-domain conventions including units and scales, and reasoning audited at the process and executable levels. Accordingly, evaluations should complement answer-level metrics with robustness suites, step-level diagnostics, and executable checks targeted to the failure modes most relevant to the application domain. We revisit these observations in Section 6 and connect them to the task-specific failure modes summarized in Section 2.

Task	Representative System	PAR	Highest APE	Primary Benchmarks	Executable Interface	Representative Performance
Geometry	GEOS	Perception + Alignment	Executable	GEOS (official & practice SAT geometry)	Equation solver over parsed text + diagram	49% accuracy on official SAT geometry questions and 61% on practice questions; on the ∼51% of questions the system chooses to answer, accuracy exceeds 96%.
Geometry	NGS / GeoQA program-supervised	Alignment + Reasoning	Executable	GeoQA, GeoQA+	Program executor over symbolic programs	60.0% accuracy on GeoQA; the improved DPE-NGS reaches 62.65% on GeoQA and 66.09% on GeoQA+.
Geometry	Inter-GPS	Alignment + Reasoning	Executable	Geometry3K, GeoQA, GEOS	Geometry DSL / theorem rules	78.3% accuracy on Geometry3K and 68.0% on GeoQA, clearly improving over earlier NGS (60% on GeoQA); also outperforms GEOS on the GEOS dataset.
Geometry	PGPSNet	Alignment + Reasoning	Executable	Geometry3K, GeoQA	Program-supervised geometry DSL	77.9% accuracy on Geometry3K and 70.4% on GeoQA.
Geometry	LANS	Alignment + Reasoning	Executable	Geometry3K, GeoQA	Geometry DSL with learned abstraction	82.3% accuracy on Geometry3K and 74.0% on GeoQA, ranking among the strongest traditional GPS systems.
Geometry	FormalGeo-style provers / FGeo-HyperGNet	Alignment + Reasoning	Executable	FormalGeo7K	FormalGeo DSL + symbolic engine	Around 85.5% overall accuracy and 87.7% step-wise accuracy on FormalGeo7K, significantly outperforming previous geometry solvers.
Geometry	GeoDRL	Alignment + Reasoning	Executable	GeoQA	RL-guided theorem selection with symbolic solver	About 89.4% accuracy on GeoQA, one of the highest reported results on this benchmark.
Geometry	Sufli-GPSC / FGeo-DRL series	Alignment + Reasoning	Executable	GeoQA, GeoQA+	RL-guided formal solver	Sufli-GPSC achieves 87.4% accuracy on GeoQA; FGeo-DRL reports 86.4% on GeoQA, offering a trade-off between peak accuracy and proof interpretability.
Geometry	E-GPS	Perception + Alignment + Reasoning	Executable	Geometry3K, GeoQA	Top-down solver + bottom-up problem generator	Reports accuracy on Geometry3K and GeoQA comparable to Inter-GPS and GeoDRL, while substantially reducing average reasoning steps and improving interpretability (exact numbers are given in the original tables).
Geometry	Pi-GPS	Perception + Alignment + Reasoning	Executable	Geo170K, Geometry3K	Large-scale GPS pipeline with geometry DSL	Claims nearly a 10-point absolute improvement over previous neuro-symbolic GPS methods on Geometry3K and maintains state-of-the-art performance on the large-scale Geo170K corpus (exact percentages reported in the original paper).
Geometry	AlphaGeometry	Reasoning	Executable	IMO-AG (Olympiad geometry)	Formal theorem prover (DDAR)	Solves 25 of 30 recent IMO-AG geometry problems; the later AlphaGeometry2 variant solves 42 of 50 problems from 2000–2024 (84% solve rate), surpassing the average human gold-medalist performance.
Charts & Tables	VisionTaPas	Alignment + Reasoning	Answer	ChartQA-H / ChartQA-M	Text + table encoder (non-pixel)	About 45.5% overall accuracy on the original ChartQA test set, with 28.72% on the harder ChartQA-H split and 53.84% on ChartQA-M.
Charts & Tables	Pix2Struct-Large	Perception + Alignment	Answer	ChartQA, AI2D, etc.	Fully visual encoder–decoder (no explicit table interface)	Achieves 58.6% relaxed accuracy on ChartQA, improving the previous VisionTaPas result from 45.5% to 58.6%.
Charts & Tables	ChartLlama	Perception + Alignment + Reasoning	Answer	ChartQA, chart-to-text, chart extraction	LLaVA-style VLM with chart-specific pre-training	Obtains 48.96% accuracy on the original ChartQA test set and 90.36% on the authors' "special charts," for an average of 69.66% across their two splits.
Charts & Tables	ChartVLM	Perception + Alignment + Reasoning	Answer	ChartX (ChartQA-like multi-task benchmark)	Chart-specialized VLM	Around 40.71% accuracy on the ChartQA-style task in ChartX, substantially outperforming general-purpose LMMs such as GPT-4V on this benchmark.
Charts & Tables	GPT-4V / GPT-4o / LLaVA	Perception + Reasoning	Answer	ChartQA, ChartInsights	—	GPT-4V/4o generally outperform open-source models such as InstructBLIP and LLaVA on chart reasoning; on the ChartInsights benchmark, GPT-4o reaches about 69.2% accuracy, whereas the mean accuracy of 19 other open and closed models is only ∼39.8%.
VWP & Mixed	LLaVA-13B	Perception + Reasoning	Answer	MathVista test	—	Achieves 25.4% overall accuracy on MathVista test, only modestly above the random baseline of 17.9%.
VWP & Mixed	CoT / PoT GPT-4 (caption + OCR tools)	Reasoning (tool-augmented)	Answer	MathVista test	External tools (image captioning + OCR)	CoT GPT-4 reaches 30.50% accuracy and PoT GPT-4 reaches 31.74% on MathVista, showing moderate gains from tool-augmented text-only pipelines.
VWP & Mixed	GPT-4V	Perception + Reasoning	Answer	MathVista test	Direct image input	Achieves 49.9% overall accuracy on MathVista test, about 15.1 points higher than Bard and still roughly 10.4 points below human performance (60.3%).
VWP & Mixed	Math-LLaVA-13B	Perception + Reasoning	Answer	MathVista testmini, MathVerse, etc.	—	Reaches 46.6% accuracy on MathVista testmini, improving over the LLaVA-1.5-13B base model by 19 absolute points and approaching GPT-4V on this split; also achieves competitive results on Math-V and related benchmarks.

Table 2: Representative systems and reported performance on shared benchmarks.