M2E-X: GROUNDING MULTI-VIEW MATHEMATICAL REASONING WITH STRUCTURED EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view mathematical problems distribute task-relevant information across several related visual inputs, requiring a reasoner to recover local symbols, mathematical entities, relations, and cross-view correspondences before solving the problem. We present M2E-X, a two-stage framework that grounds downstream multi-view mathematical reasoning in explicit structured evidence. Stage 1 performs answer-free multi-view mathematical perception and constructs a structured mathematical scene representation, $\mathcal{G}_{\mathrm{math}}$, containing grounded entities, text, typed intra-view and cross-view relations, entity bindings, shared scene memory, and evidence-selection signals. Stage 2 compiles task-specific evidence from $\mathcal{G}_{\mathrm{math}}$ through sparse proof-evidence retrieval, hierarchical question alignment, and graph counterfactual verification with adaptive retrieval control, yielding a task-specific subgraph $\mathcal{P}_q$. The selected evidence is converted into typed entity, relation, binding, and measurement atoms and deterministically serialized as structured mathematical evidence $\mathcal{E}_q$ for an otherwise frozen multimodal reasoner. Neither stage is trained to generate the final answer, and the downstream evaluation keeps the reasoner fixed. We evaluate the framework with stage-wise evidence diagnostics and a controlled paired RAW/FULL protocol on sealed multi-view mathematical benchmarks, where RAW and FULL differ only by the addition of model-generated evidence.

1 INTRODUCTION

Multimodal mathematical reasoning requires more than recognizing text or objects in an image. A solver must identify mathematical entities, recover measurements and symbolic labels, connect spatial relations to the question, and then use the resulting evidence correctly. Recent benchmarks have made these requirements increasingly explicit. MathVista evaluates broad mathematical reasoning in visual contexts (Lu et al., 2024), MathVerse probes whether multimodal models actually use diagram information (Zhang et al., 2024), and MATH-Vision emphasizes competition-level visual mathematics (Wang et al., 2024). These benchmarks are primarily organized around a single visual context. MV-MATH instead targets multi-visual mathematical reasoning with 2,009 problems whose questions may depend on several interleaved images (Wang et al., 2025a). More recent benchmarks such as OMIBench and VC-Bench further stress multi-image reasoning with evidence distributed across visual inputs (Chen et al., 2026; Wang et al., 2025b).

Multi-view problems create a specific structural burden. A relevant measurement may appear only in a detail view, the geometric structure needed to interpret it may appear in a global view, and the same mathematical entity may be represented differently across projections, states, options, or complementary diagrams. No single image is guaranteed to contain a sufficient description of the problem. A downstream multimodal reasoner must therefore infer which visual regions matter, associate related entities across views, and preserve the resulting mathematical relations while also performing the final reasoning. MV-MATH reports substantial difficulty for existing multimodal models in this setting (Wang et al., 2025a). We treat this difficulty as a motivation for making task-relevant mathematical evidence explicit rather than requiring the final reasoner to reconstruct all of it implicitly.

Our approach is M2E-X, a two-stage structured-evidence framework for multi-view mathematical reasoning. Given a question q and visual inputs $\mathcal{I} = \{I_m\}_{m=1}^{M}$, Stage 1 constructs a reusable *struc-*

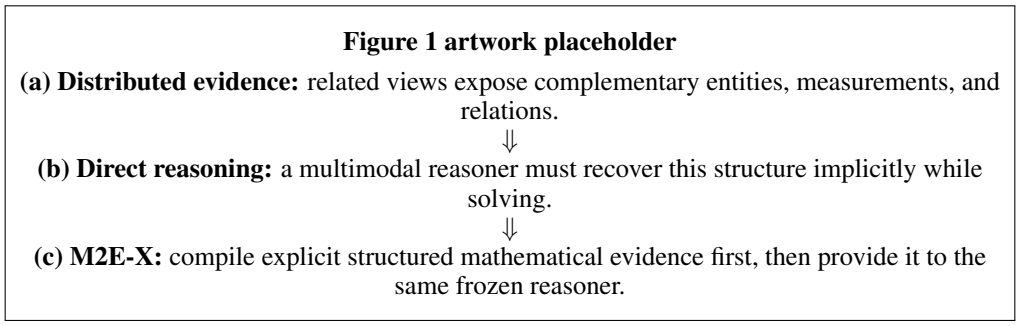


Figure 1: Motivation and core idea of M2E-X. Multi-view mathematical reasoning can require complementary evidence that is distributed across related visual inputs. M2E-X first constructs and verifies task-relevant mathematical evidence, then supplies that evidence to a frozen downstream reasoner. The final artwork will use a real benchmark example and will not contain synthetic performance claims.

tured mathematical scene representation $\mathcal{G}_{\mathrm{math}}$ without access to final answers or solution-derived supervision. Stage 2 conditions on the question and compiles a sparse, task-specific evidence subgraph $\mathcal{P}_q$ from $\mathcal{G}_{\mathrm{math}}$. A verifier estimates whether the selected evidence is sufficient and structurally consistent; if not, retrieval is expanded within a bounded evidence budget. The verified selection is converted into typed mathematical evidence atoms and deterministically serialized into $\mathcal{E}_q$. A frozen multimodal reasoner then receives the original problem, with $\mathcal{E}_q$ added only in the FULL condition.

This design separates structured evidence construction from downstream answer generation. The separation serves two purposes. First, the perception and evidence pipeline can be evaluated directly through grounding, relation, retrieval, coverage, and verification diagnostics rather than only through final-answer accuracy. Second, the same evidence interface can be evaluated with different frozen downstream reasoners, avoiding task-specific reasoner fine-tuning as a confound in the central RAW/FULL comparison.

Our technical contributions are:

- **Answer-free structured mathematical scene modeling.** We construct $\mathcal{G}_{\mathrm{math}}$, a reusable multi-view representation that combines grounded mathematical entities, byte-level text, typed intra-view and cross-view relations, explicit entity bindings, shared scene memory, and question-conditioned evidence-selection signals, while keeping per-view entity parsing independent of final-answer generation.
- **Task-conditioned structured evidence compilation.** We compile $\mathcal{G}_{\mathrm{math}}$ into a task-specific proof-evidence subgraph $\mathcal{P}_q$ with Sparse Cross-View Proof-Evidence Retrieval (SCVR), Hierarchical Dynamic Task Alignment (DTA), and Graph Counterfactual Verification and Retrieval Control (GCVR), coupled through a verify-and-expand retrieval loop.
- **A typed mathematical evidence interface.** We convert selected neural evidence into explicit entity, relation, binding, and measurement atoms and deterministically serialize them as $\mathcal{E}_q$ for frozen multimodal reasoners. This interface exposes what evidence is supplied to the reasoner without generating a solution trace or final answer.

2 Related Work

2.1 Multimodal and multi-view mathematical reasoning

Visual-mathematical benchmarks have progressed from specialized diagram and chart tasks toward general multimodal reasoning. PlotQA evaluates reasoning over scientific plots and includes a pipeline that reconstructs structured tables from visual elements before question answering (Methani et al., 2020). GeoQA studies multimodal numerical reasoning over geometry diagrams and text (Chen et al., 2021). More recent benchmarks target general-purpose multimodal models: MathVista spans diverse mathematical and visual tasks (Lu et al., 2024), MathVerse controls how

much problem information is available in text versus diagrams (Zhang et al., 2024), and MATH-Vision collects competition problems with visual contexts (Wang et al., 2024).

Our target differs in requiring multiple related visual inputs. MV-MATH formalizes multi-visual mathematical reasoning and includes multiple-choice, free-form, and multi-step problems (Wang et al., 2025a). OMIBench evaluates Olympiad-level reasoning when required evidence is distributed across multiple images, including a mathematics subset (Chen et al., 2026). VC-Bench studies explicit visual dependency in multi-image mathematical reasoning (Wang et al., 2025b). These works establish evaluation settings for multi-image reasoning. M2E-X is instead an evidence-construction method: it is trained without these downstream benchmarks and is evaluated on them as a front-end to frozen reasoners.

2.2 Structured mathematical perception and symbolic interfaces

A separate line of work explicitly parses mathematical visuals before solving. Inter-GPS maps geometry text and diagrams to a formal language and invokes symbolic theorem reasoning (Lu et al., 2021). PGDPNet focuses on fine-grained plane-geometry diagram parsing and relation recovery, together with the PGDP5K dataset (Zhang et al., 2022). GeoQA likewise uses program supervision for geometry reasoning (Chen et al., 2021). These systems motivate explicit representations of mathematical content, but they are primarily designed for specific geometry or chart-solving pipelines.

M2E-X retains the benefit of explicit structure while targeting heterogeneous multi-view inputs and a frozen general-purpose multimodal reasoner. Its intermediate representation is not a theorem program and Stage 2 is not a symbolic solver. Instead, $\mathcal{G}_{\mathrm{math}}$ captures observable mathematical scene structure, and $\mathcal{E}_q$ exposes selected evidence that remains agnostic to the downstream answer generator.

2.3 Intermediate evidence for visual reasoning

Several visual-reasoning approaches externalize intermediate computation. Visual CoT trains multimodal models with intermediate visual regions and reasoning steps (Shao et al., 2024). VisProg and ViperGPT decompose visual tasks into executable programs that call external vision or language modules (Gupta & Kembhavi, 2023; Surís et al., 2023). These approaches show the value of separating intermediate visual processing from final task execution. M2E-X takes a narrower evidence-centered route for mathematics: it does not generate chain-of-thought, does not synthesize an executable reasoning program, and does not retrieve external knowledge. It compiles task-relevant evidence only from the problem's own visual inputs and supplies that evidence through a deterministic typed interface.

3 Method

3.1 Overview

Let q denote a mathematical question and $\mathcal{I} = \{I_m\}_{m=1}^{M}$ its related visual inputs. M2E-X factorizes evidence construction into

$$(q, \mathcal{I}) \xrightarrow{\text{Stage 1}} \mathcal{G}_{\mathrm{math}} \xrightarrow{\text{Stage 2}} \mathcal{P}_q \xrightarrow{\text{atomization/serialization}} \mathcal{E}_q. \tag{1}$$

Stage 1 performs answer-free multi-view mathematical perception. Stage 2 retrieves, aligns, and verifies task-specific evidence while preserving the identities of selected entities. The final evidence $\mathcal{E}_q$ is appended to the original problem only at downstream inference time.

3.2 Answer-Free Multi-View Mathematical Perception

Global-local visual encoding. The canonical Stage 1 configuration uses a DINOv2 ViT-B/14 global visual encoder (Oquab et al., 2024) together with a ResNet50 (He et al., 2016) multi-scale local branch. The local branch aggregates features from multiple residual stages, projects them to the global feature width, aligns them to the ViT patch grid, and fuses them with the global tokens through a learned gate. Full-grid tokens are retained before entity parsing so that small labels and

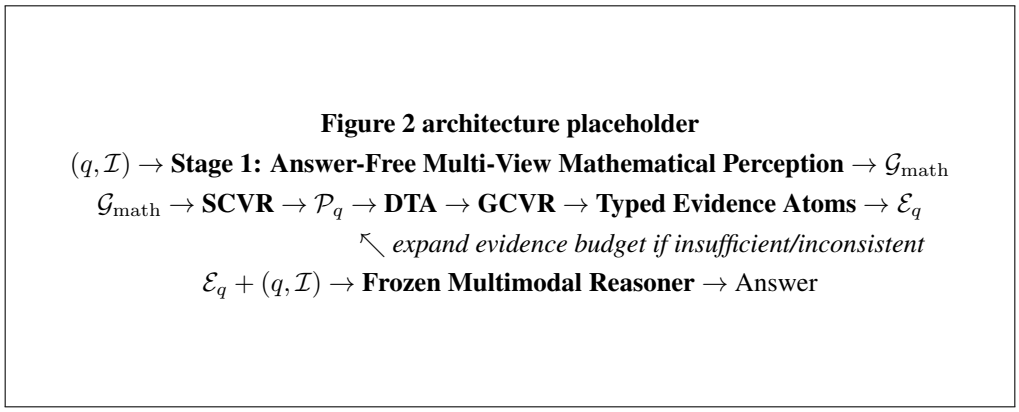


Figure 2: M2E-X: from multi-view perception to verified structured evidence. Stage 1 constructs $\mathcal{G}_{\text{math}}$ without final-answer supervision. Stage 2 retrieves a sparse task-specific subgraph, aligns it with the question and scene context, and verifies sufficiency and consistency through an adaptive verify-and-expand loop. The final answer is produced only by the frozen downstream reasoner.

line markings are not discarded by early token sampling. The same visual encoder is reused for selected ROI crops during active re-observation.

Mathematical entity parsing. For each valid view, a fixed-capacity set decoder predicts entity presence, entity type, normalized bounding box, byte-level text, and a view-role label. The set-prediction formulation follows the general bipartite-matching principle used by DETR-style decoders (Carion et al., 2020), but the prediction vocabulary is specialized to mathematical entities and relations. Crucially, per-view entity parsing is image-only: the question does not change the entity labels, text predictions, or boxes. This preserves the interpretation of Stage 1 as mathematical perception rather than answer-conditioned detection.

Let $e_{m,i} \in \mathbb{R}^d$ denote the ith entity feature in view m. For each entity the model predicts

$$p^{\text{obj}}_{m,i}, \quad p^{\text{type}}_{m,i}, \quad b_{m,i}, \quad p^{\text{text}}_{m,i}. \tag{2}$$

Training uses set matching and supervision-availability masks, so an annotation that is absent in a public source is not converted into a negative label.

Multi-view relational structuring and scene memory. Entity features are organized into a typed graph. Separate heads predict intra-view relations, cross-view relations, and cross-view entity binding. Sequence position is introduced only for samples in which view order is semantically meaningful, such as state transformations; unordered image sets remain view-order neutral during per-view parsing. A graph message-passing layer updates entity features, and a bank of learned memory queries reads from all valid entities to form shared scene-memory tokens. This yields a representation that exposes both local evidence and cross-view structure.

Question-conditioned active evidence perception. The question is encoded into token-level features and a pooled query feature. These features are used by an evidence controller, not by the image-only entity parser. The controller scores critical views and regions, predicts an evidence order and stopping distribution, and selects a bounded set of ROI crops for re-observation. Selected crops are re-encoded from the original pixels and used to refine entity features before the relational graph, scene memory, and evidence-control outputs are recomputed. This implements question-conditioned active perception without turning Stage 1 into a final reasoning module.

The resulting structured scene contract is denoted by

$$\mathcal{G}_{\text{math}} = \big(E,\, B,\, T,\, R^{\text{intra}},\, R^{\text{cross}},\, C,\, H,\, Q,\, S\big)\,, \tag{3}$$

where E denotes entity features, B boxes and objectness, T entity text/type predictions, R^{intra} and R^{cross} typed relations, C cross-view bindings, H shared scene memory, Q token-level and pooled question features, and S evidence-selection signals. The implementation passes these tensors to Stage 2 through a fixed contract; it does not expose final-answer or solution fields.

3.3 TASK-CONDITIONED STRUCTURED EVIDENCE COMPILATION

Sparse Cross-View Proof-Evidence Retrieval (SCVR). Stage 2 begins from the predicted entity graph in $\mathcal{G}_{\text{math}}$. Here, *proof-evidence* denotes evidence that is relevant to solving the current problem; it does not denote a generated formal theorem proof. SCVR computes a question-conditioned seed score for each candidate entity. For pooled query feature q_f and entity feature e_i, the semantic seed scorer receives

$$[q_f;\, e_i;\, q_f \odot e_i;\, |q_f - e_i|], \tag{4}$$

and combines the resulting score with an objectness prior. Typed intra-view, cross-view, and binding predictions define a weighted graph. SCVR selects a small seed set and expands it over high-scoring graph edges subject to an evidence budget K and a bounded number of hops. The output preserves selected entity IDs, view/slot IDs, typed edges, retrieval scores, and retrieval source, forming a sparse task-specific subgraph $\mathcal{P}_q$.

Hierarchical Dynamic Task Alignment (DTA). SCVR decides *which* evidence items to retain; DTA conditions these items on the question and scene context while preserving their identities. It applies three alignment levels. First, entity-level query alignment uses cross-attention from selected entities to token-level question features. Second, relation-level structural alignment performs typed message passing over the selected intra-view, cross-view, and binding edges. Third, scene-level context alignment attends from the selected evidence to Stage 1 memory tokens. A learned hierarchy gate combines the entity-, relation-, and scene-level representations for each selected item. Unlike a latent bottleneck that collapses the selection to one vector, DTA returns one aligned representation for each retained evidence item.

Graph Counterfactual Verification and Retrieval Control (GCVR). GCVR receives the DTA-aligned evidence together with the explicit selected graph. It predicts two logits,

$$z_{\text{suf}} = h_{\text{suf}}(\mathcal{P}_q, q_f), \qquad z_{\text{con}} = h_{\text{con}}(\mathcal{P}_q, q_f), \tag{5}$$

for evidence sufficiency and structural consistency. Thresholded decisions determine whether the current evidence set is ready. If either criterion fails, the evidence budget is increased and SCVR–DTA–GCVR is run again:

$$K_{t+1} = \begin{cases} K_t, & \text{if sufficient and consistent,} \\ \min(K_t + \Delta K, K_{\max}), & \text{otherwise.} \end{cases} \tag{6}$$

The loop terminates when both criteria are satisfied or the maximum budget is reached. GCVR is trained with retrieval-derived verifier targets and synthetic structural counterfactuals. A critical-evidence deletion is supervised as insufficient only when the proof generator guarantees loss of sufficiency; a hard binding swap is supervised as inconsistent only when the generator guarantees an incorrect cross-view correspondence. Unlabeled counterfactual dimensions are not treated as negatives.

3.4 STRUCTURED MATHEMATICAL EVIDENCE

The selected neural evidence is converted into a deterministic intermediate representation consisting of four atom types:

- **Entity atoms** record selected entity identity, type, label, location, view/slot provenance, and retrieval metadata.
- **Relation atoms** encode typed intra-view or cross-view predicates between selected entities.
- **Binding atoms** record explicit cross-view entity correspondence.
- **Measurement atoms** normalize a conservative subset of numeric, fractional, and unit-bearing measurements linked to their target entities.

The compiler contains no trainable parameters and does not consume answers, solutions, raw visual features, scene-memory tokens, or question features. It only atomizes the selected Stage 1/SCVR outputs.

Figure 3 result-dependent placeholder

Real multi-view benchmark inputs → predicted Stage 1 entities/relations → SCVR-selected subgraph → typed evidence atoms → deterministic evidence text.

A real model-generated example will be inserted after sealed inference.

Figure 3: From multi-view inputs to structured mathematical evidence. The final figure will trace one real evaluation sample from the original views through the predicted scene structure, selected proof-evidence subgraph, typed atoms, and serialized evidence. No ground-truth evidence or fabricated model output will be used in this figure.

A deterministic serializer then converts the atom bundle into $\mathcal{E}_q$ using compact entity, measurement, relation, and binding records. For downstream FULL evaluation, the serializer outputs evidence only; it does not repeat the question or add a new solving instruction. This design makes the experimental difference between RAW and FULL explicit.

3.5 Evidence-Grounded Frozen Reasoning

Let F_θ denote a frozen multimodal reasoner. The controlled downstream comparison uses

$$\mathrm{RAW}(q, \mathcal{I}) = F_\theta(q, \mathcal{I}), \tag{7}$$

$$\mathrm{FULL}(q, \mathcal{I}) = F_\theta(q, \mathcal{I}, \mathcal{E}_q). \tag{8}$$

The reasoner parameters θ, original images, question, decoding configuration, and scoring procedure are held fixed. The intended experimental intervention is therefore the presence of model-generated structured evidence $\mathcal{E}_q$. We do not fine-tune the downstream reasoner on MV-MATH or on M2E-X evidence.

4 Data and Experimental Setup

4.1 Stage 1 perception data

Stage 1 combines public mathematical-visual data with an internally generated multi-view perception corpus. The public sources are PGDP5K (Zhang et al., 2022), AutoGeo-100K (Huang et al., 2024), PlotQA (Methani et al., 2020), and filtered `mathQa` samples from the National Smart Education Platform data-innovation competition (mat, 2024). The sources provide different annotation types, so Stage 1 uses explicit supervision-availability masks rather than assuming every sample has entity, relation, binding, or active-evidence labels.

The `mathQa` preprocessing deserves separate mention because the original competition data can associate more than one image with a problem. We retain the question-stem image (`_0.png`) and discard later explanation/solution-step images (`_1+`) to prevent solution-derived visual leakage. The resulting Stage 1 subset is therefore predominantly single-image and supplies only the labels supported by the adapter.

`M2EX-MultiView-v1` is a fully programmatic synthetic perception corpus created by the project. Each sample starts from an internal latent mathematical scene, derives related views and typed relations, and renders actual PNG images with programmatic drawing routines. It does not copy external questions, external image assets, or downstream benchmark samples. The corpus covers ten scene families spanning geometry, transformations, projections, global/detail relationships, option sets, and redundant/conflicting evidence. Its role is to supervise multi-view entities, relations, bindings, view roles, and active-evidence signals that are not uniformly available in the public sources.

The Stage 1 schema is answer-free by construction. It rejects answer, solution, rationale, and chain-of-thought fields, and the downstream MV-MATH benchmark is excluded from Stage 1 supervision.

4.2 Stage 2 proof-evidence supervision

Stage 2 is trained on `M2EX-Proof-v1`, a fully synthetic proof/schema supervision corpus generated from internal templates. The formal split contains 10,000 training, 1,000 validation, and 1,000 test rows. Rows contain scene entities and relations, proof node IDs, proof edges, bridge edges, sufficiency/necessity metadata, and labeled structural counterfactuals. Train, validation, and test use split-specific proof templates and manifest fingerprints. Stage 2 optimization consumes proof-evidence selection, alignment, verification, and counterfactual targets only; it is not trained to generate final answers or chain-of-thought.

Grounded Stage 2 targets are mapped to frozen Stage 1 predicted slots. A matched predicted entity is accepted for proof supervision only when its objectness probability is at least 0.5 and its box IoU with the target is at least 0.5. Unmappable proof items remain end-to-end misses rather than being silently converted to negative Stage 2 labels.

4.3 Downstream benchmarks and frozen reasoners

MV-MATH (Wang et al., 2025a) is the primary sealed downstream benchmark. It contains 2,009 multi-visual mathematical problems and is never used for Stage 1 or Stage 2 training. The final study is designed to include at least two additional external multi-image evaluations, with the mathematics subset of OMIBench (Chen et al., 2026) and VC-Bench (Wang et al., 2025b) as the current priority. MATH-Vision (Wang et al., 2024) is reserved as an optional single-image control to test whether any benefit is specific to distributed multi-view evidence rather than a generic prompt-augmentation effect.

We evaluate frozen multimodal reasoners from multiple model families under matched RAW/FULL conditions. For each reasoner, the checkpoint, prompt structure, decoding configuration, and scoring procedure are held fixed across the two conditions.

4.4 Metrics and statistical protocol

Stage 1 is evaluated with entity type accuracy, mean box IoU, OCR exact match, view-role accuracy, intra-view and cross-view relation macro-F1, binding precision/recall/F1, critical-view recall@K, critical-region localization metrics, and stopping accuracy. These metrics respect annotation-availability masks. Relation denominators include edges incident to missed gold entities, while false object predictions can contribute false relation predictions.

Stage 2 reports conditional retrieval metrics over grounded/mappable proof targets, including ProofNodeRecall@K, ProofNodePrecision@K, ProofEdgeRecall@K, and ProofBridgeRecall@K. We separately report end-to-end proof coverage over the full ground-truth proof so that Stage 1 grounding misses cannot disappear from the denominator. Adaptive verification is summarized by final evidence budget and verify-and-expand rounds. GCVR system contribution is tested through ablation.

For final-answer evaluation, RAW and FULL are paired by problem ID. We report official accuracy for each condition, the paired absolute accuracy difference, a 95% confidence interval for the paired difference, and an exact two-sided McNemar test for binary correctness. We additionally report the counts of RAW-wrong/FULL-correct recoveries and RAW-correct/FULL-wrong regressions. The word "significant" is reserved for comparisons with an explicitly reported statistical test.

5 Experiments

5.1 Controlled paired RAW/FULL evaluation

Table 1 is the central downstream comparison. Each row uses one frozen reasoner under matched RAW and FULL conditions. The main quantity of interest is the within-reasoner paired change after adding $\mathcal{E}_q$, rather than an unpaired leaderboard comparison across different systems.

Table 1: Controlled paired RAW/FULL evaluation across frozen multimodal reasoners on MV-MATH. All values are intentionally left as TBD until sealed runs are complete.

Frozen reasoner	RAW acc.	FULL acc.	Δ acc. $\uparrow$	95% CI	McNemar p
TBD checkpoint 1	TBD	TBD	TBD	TBD	TBD
TBD checkpoint 2	TBD	TBD	TBD	TBD	TBD
TBD checkpoint 3	TBD	TBD	TBD	TBD	TBD
Optional checkpoint 4/5	TBD	TBD	TBD	TBD	TBD

Figure 4 result placeholder

(a) Paired outcome transitions: both correct, recovered, regressed, both wrong.
(b) Paired FULL−RAW accuracy differences with 95% confidence intervals across pre-specified subsets.

No points or bars will be drawn until official per-sample correctness is available.

Figure 4: Paired RAW/FULL outcome changes on MV-MATH. The final figure will visualize sample-level transitions and paired accuracy differences with uncertainty intervals. Recovered and regressed cases will be shown together to avoid success-only reporting.

5.2 Cross-benchmark and multi-view generalization

The primary generalization question is whether the same evidence interface transfers without downstream reasoner fine-tuning. Table 2 reserves the cross-benchmark comparison. MV-MATH remains the primary benchmark; external rows are included only after complete evaluation on the official splits or evaluation protocols. If all planned external runs cannot be completed, the main-text table will instead report the official multi-view dependency breakdown available in MV-MATH, with extended benchmark results moved to the appendix.

5.3 Stage-wise evidence quality

A final-answer gain alone would not identify where the evidence pipeline succeeds or fails. Table 3 therefore connects the three paper-level contributions to direct diagnostics. Stage 1 metrics test whether $\mathcal{G}_{\mathrm{math}}$ captures mathematical entities and cross-view structure. Stage 2 conditional metrics test retrieval given grounded evidence, whereas end-to-end coverage retains Stage 1 grounding failures. Adaptive-budget statistics test the behavior of verification-driven retrieval control rather than treating the verifier score itself as evidence quality.

5.4 Ablations and evidence controls

The ablation design follows the contribution claims rather than treating every implementation switch as equally important. Removing SCVR tests whether sparse typed graph retrieval matters; removing hierarchical DTA tests question/structure/scene alignment; disabling GCVR verify-and-expand tests adaptive verification and retrieval control; and removing counterfactual supervision tests the training signal used by the verifier. A content-matched flat-evidence control is particularly important. It should preserve as much evidence content as possible while removing the typed structured organization, testing the alternative explanation that FULL improves only because it contains additional text.

Table 2: Cross-benchmark generalization of M2E-X structured evidence. The exact external suite is finalized only after complete verified runs; MATH-Vision is an optional single-image control.

Benchmark	Setting	RAW	FULL	$\Delta \uparrow$
MV-MATH	multi-view, primary	TBD	TBD	TBD
OMIBench	mathematics subset	TBD	TBD	TBD
VC-Bench	multi-image	TBD	TBD	TBD
MATH-Vision	single-image control	TBD	TBD	TBD

Table 3: Stage-wise evidence quality and end-to-end coverage. The compact main-text table will contain the most diagnostic metrics; full per-source and per-family results will be reported in the appendix.

Stage	Metric	Result
Stage 1	Entity type accuracy	TBD
Stage 1	Mean box IoU	TBD
Stage 1	OCR exact match	TBD
Stage 1	Cross-view relation macro-F1	TBD
Stage 1	Binding recall	TBD
Stage 2	Proof mapping recall	TBD
Stage 2	ProofNodeRecall@K	TBD
Stage 2	ProofBridgeRecall@K	TBD
End-to-end	Exact proof coverage	TBD
Verification	Avg. final evidence budget / rounds	TBD

5.5 Failure analysis

We predefine failure categories before inspecting the final headline result. On proof-supervised data, automatic categories separate Stage 1 grounding failure, Stage 2 retrieval failure, structural edge/bridge retrieval failure, and verification/retrieval-control failure. For MV-MATH, every paired sample first belongs to one of four outcome groups: both correct, recovered (RAW wrong and FULL correct), regressed (RAW correct and FULL wrong), or both wrong.

For a manually inspected subset, the error codebook distinguishes perception/OCR error, cross-view correspondence error, relevant-evidence retrieval error, verification/consistency error, serialization/evidence-communication error, downstream reasoning failure despite adequate evidence, answer-extraction/evaluation-format error, and unclear cases. Recovered and regressed examples are analyzed under the same codebook. The planned annotation protocol uses two independent annotators, reports inter-annotator agreement with Cohen's κ, and adjudicates disagreements after the initial labels are frozen.

6 Limitations and Broader Impacts

M2E-X has several limitations independent of the final benchmark scores. First, Stage 2 relies on synthetic proof/schema supervision. Although the downstream benchmarks are held out from training, synthetic proof templates cannot cover the full diversity of real mathematical representations and reasoning requirements. Second, Stage 2 is bounded by Stage 1 grounding quality. OCR errors, missed entities, incorrect relation labels, or incorrect cross-view bindings can remove evidence before retrieval. Third, deterministic evidence serialization is intentionally compact and can omit visual information that a downstream reasoner might otherwise exploit directly. Fourth, any downstream benefit depends on the frozen reasoner's ability to use the supplied evidence; evaluation across multiple model families reduces but does not remove this dependence. Finally, the current method does not provide a formal guarantee that serialized evidence is correct or sufficient on real benchmark instances.

Structured evidence may be useful for tutoring, solution auditing, and systems that need explicit intermediate mathematical observations. The same capability can also facilitate automated answering in educational assessment settings. Producing evidence rather than answers does not eliminate this

Table 4: Ablations and evidence controls. Downstream and evidence-pipeline metrics are reported together so that an ablation is tied to the mechanism it is intended to test.

Variant	MV-MATH acc.	Δ vs. RAW	E2E proof coverage	Avg. budget
M2E-X (full)	TBD	TBD	TBD	TBD
w/o SCVR graph expansion	TBD	TBD	TBD	TBD
w/o hierarchical DTA	TBD	TBD	TBD	TBD
w/o GCVR verify-and-expand	TBD	TBD	TBD	TBD
w/o counterfactual supervision	TBD	TBD	TBD	TBD
Content-matched flat evidence	TBD	TBD	–	–

risk because the downstream reasoner still produces the final answer. Incorrect but plausible evidence can additionally mislead users. Applications in education should therefore distinguish assistance from assessment, retain access to original visual inputs, and expose uncertainty or verification diagnostics where appropriate.

7 Conclusion

We introduced M2E-X, a two-stage framework that grounds multi-view mathematical reasoning in structured evidence before final answer generation. Stage 1 constructs an answer-free structured mathematical scene representation $\mathcal{G}_{\mathrm{math}}$; Stage 2 retrieves, aligns, and verifies a task-specific subgraph $\mathcal{P}_q$; and a deterministic typed interface converts the selected evidence into $\mathcal{E}_q$ for a frozen multimodal reasoner. The design makes perception, cross-view structure, evidence retrieval, verification, and downstream reasoning separately measurable. The final empirical conclusions will be limited to the sealed RAW/FULL comparisons and component analyses supported by the completed experiments.

AI use statement

Working-draft placeholder. The final submission will disclose the use of generative AI tools in accordance with the ICLR 2027 author policy, including the specific roles of such tools in literature search, code assistance, and manuscript drafting or editing where applicable. All citations, implementation claims, and experimental results will be independently checked by the authors, who remain responsible for the final submission.

Ethics statement

The work does not introduce human-subject experiments. The internally generated `M2EX-MultiView-v1` and `M2EX-Proof-v1` corpora use programmatic scenes and templates rather than copied external questions or images. Public datasets are used only for their documented research purposes; exact license and redistribution statements will be verified and reported before submission. The main application risks are educational misuse and over-reliance on incorrect model-generated evidence, as discussed in Section 6.

Reproducibility statement

The implementation separates Stage 1, Stage 2, and downstream evaluation and binds formal runs to configuration, checkpoint, dataset/split, and prediction artifacts using manifests and SHA-256 identifiers. The appendix documents data provenance, supervision boundaries, metric definitions, and the planned failure-analysis protocol. Final submission materials will additionally report the exact frozen reasoner checkpoints, decoding settings, hyperparameters, compute environment, and all paper-eligible result artifacts needed to reproduce the central RAW/FULL comparison.

References

National smart education platform data innovation application activity: Multimodal question knowledge-point tag prediction. Baidu AI Studio competition page, 2024. URL `https://`

aistudio.baidu.com/competition/detail/1208/0/task-definition. Accessed 2026-08-26.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020. doi: 10.1007/978-3-030-58452-8_13. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/832_ECCV_2020_paper.php.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 513–523, 2021. doi: 10.18653/v1/2021.findings-acl.46. URL https://aclanthology.org/2021.findings-acl.46/.

Qiguang Chen, Chengyu Luan, Jiajun Wu, Qiming Yu, Yi Yang, Yizhuo Li, Jingqi Tong, Xiachong Feng, Libo Qin, and Wanxiang Che. OMIBench: Benchmarking olympiad-level multi-image reasoning in large vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45100–45135, 2026. doi: 10.18653/v1/2026.acl-long.2090. URL https://aclanthology.org/2026.acl-long.2090/.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962, 2023. doi: 10.1109/CVPR52729.2023.01436. URL https://openaccess.thecvf.com/content/CVPR2023/html/Gupta_Visual_Programming_Compositional_Visual_Reasoning_Without_Training_CVPR_2023_paper.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Zihan Huang, Tao Wu, Wang Lin, Shengyu Zhang, Jingyuan Chen, and Fei Wu. Autogeo: Automating geometric image dataset creation for enhanced geometry understanding. *arXiv preprint arXiv:2409.09039*, 2024. URL https://arxiv.org/abs/2409.09039.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 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)*, pp. 6774–6786, 2021. doi: 10.18653/v1/2021.acl-long.528. URL https://aclanthology.org/2021.acl-long.528/.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/663bce02a0050c4a11f1eb8a7f1429d3-Abstract-Conference.html.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1527–1536, 2020. URL https://openaccess.thecvf.com/content_WACV_2020/html/Methani_PlotQA_Reasoning_over_Scientific_Plots_WACV_2020_paper.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien

Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8612–8642, 2024. doi: 10.52202/079017-0275. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0ff38d72a2e0aa6dbe42de83a17b2223-Abstract-Datasets_and_Benchmarks_Track.html.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11888–11898, 2023. doi: 10.1109/ICCV51070.2023.01092. URL https://openaccess.thecvf.com/content/ICCV2023/html/Suris_ViperGPT_Visual_Inference_via_Python_Execution_for_Reasoning_ICCV_2023_paper.html.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with the MATH-Vision dataset. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, 2024. doi: 10.52202/079017-3014. URL https://proceedings.nips.cc/paper_files/paper/2024/hash/ad0edc7d5fa1a783f063646968b7315b-Abstract-Datasets_and_Benchmarks_Track.html.

Peijie Wang, Zhong-Zhi Li, Fei Yin, Dekang Ran, and Cheng-Lin Liu. MV-MATH: Evaluating multimodal math reasoning in multi-visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19541–19551, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_MV-MATH_Evaluating_Multimodal_Math_Reasoning_in_Multi-Visual_Contexts_CVPR_2025_paper.html.

Zhikai Wang, Jiashuo Sun, Wenqi Zhang, Zhiqiang Hu, Xin Li, Fan Wang, and Deli Zhao. Benchmarking multimodal mathematical reasoning with explicit visual dependency. *arXiv preprint arXiv:2504.18589*, 2025b. URL https://arxiv.org/abs/2504.18589.

Ming-Liang Zhang, Fei Yin, Yi-Han Hao, and Cheng-Lin Liu. Plane geometry diagram parsing. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, 2022. URL https://www.ijcai.org/proceedings/2022/0228.pdf.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision (ECCV)*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/1270_ECCV_2024_paper.php.

A Data Provenance and Supervision Boundaries

A.1 Stage 1 source roles

The Stage 1 mixture intentionally combines heterogeneous sources. PGDP5K contributes fine-grained plane-geometry diagram structure (Zhang et al., 2022). AutoGeo-100K supplies automatically generated geometry image–text pairs (Huang et al., 2024). PlotQA supplies plot-oriented visual structure (Methani et al., 2020). Filtered `mathQa` supplies weak view-role supervision from real multimodal mathematics questions (mat, 2024). `M2EX-MultiView-v1` supplies fully controlled multi-view relations, bindings, view roles, critical evidence, and active-evidence supervision. The canonical target builder carries an explicit supervision mask for each annotation family, preventing missing public-source labels from becoming implicit negative examples.

A.2 M2EX-MULTIVIEW-V1

M2EX-MultiView-v1 is generated from internal latent scenes and programmatically rendered PNGs. The formal family taxonomy is:

> plane_geometry, circle_geometry, coordinate, transformation, fold_cut_unfold, solid_projection, global_detail, complementary, option_set, redundant_conflict.

The recommended corpus size is 2,400 training, 300 validation, and 300 test samples at 336×336 resolution. Each split uses a split-specific seed. Positive binding supervision is derived from explicit same-entity relations rather than canonical-ID coincidence. Generator audits include image existence, geometry/text overlap checks, critical-region constraints, multi-view utility checks, and diversity caps.

A.3 M2EX-PROOF-V1

M2EX-Proof-v1 uses the same ten broad families but a different generator and supervision target. It contains internal proof templates with explicit proof nodes, proof edges, cross-view bridge edges, and counterfactual labels. The formal split is 10,000/1,000/1,000 for train/validation/test, and the three splits use distinct proof-template sets. Formal manifests record row counts, JSONL SHA-256 values, and sample-ID SHA-256 values. The Stage 2 evaluator binds paper-eligible evaluation to these manifests.

The generator can store a synthetic answer for internal template consistency, but Stage 2 losses do not train answer generation. The Stage 2 objective contains proof-node selection, bridge retrieval, question-evidence alignment, verification, and counterfactual terms only.

B METRIC DEFINITIONS

Grounded Stage 1 metrics. Matched entity predictions are considered grounded under the metric-specific objectness and IoU gates. Entity type and OCR metrics are computed only where the corresponding annotations are supervised. For relation metrics, missed entities retain their incident gold relations in the denominator, and predicted edges involving false object detections count against precision.

Conditional Stage 2 retrieval. Conditional proof metrics evaluate Stage 2 only on proof items that can be mapped to grounded Stage 1 predictions. They answer whether Stage 2 retrieves the correct evidence *given* that Stage 1 produced a mappable entity.

End-to-end proof coverage. End-to-end coverage uses the full generated proof as denominator, including proof items missed during Stage 1 grounding. This metric therefore measures the full perception-to-retrieval pipeline rather than Stage 2 in isolation.

Paired downstream statistics. For each reasoner, RAW and FULL predictions are joined by benchmark problem ID. The paired accuracy difference is computed from per-problem correctness. Confidence intervals are obtained by paired bootstrap resampling over problem IDs. McNemar's test uses only the discordant pairs, i.e., recovered and regressed samples.

C FAILURE ANALYSIS CODEBOOK

The automatic pipeline categories are:

1. **Grounding failure:** required proof evidence cannot be mapped to Stage 1 predictions under the grounding gate.
2. **Retrieval failure:** required evidence is mappable but not selected by SCVR.
3. **Structural retrieval failure:** required nodes are selected but a necessary proof or bridge edge is absent.

4. **Verification/control failure:** the selected evidence remains insufficient or inconsistent and the retrieval-control behavior does not recover the missing structure within the allowed budget.

The downstream manual codebook is:

> P: perception/OCR; C: cross-view correspondence; R: evidence retrieval; V: verification/consistency; S: serialization/evidence communication; D: downstream reasoning despite adequate evidence; E: answer extraction/evaluation format; U: unclear.

No new category will be introduced after outcome inspection without documenting the change.

D HYPERPARAMETERS AND COMPUTE

Table 5: Core architecture hyperparameters from the canonical implementation. Training and hardware fields that depend on final formal runs remain TBD.

Item	Value
Stage 1 image size	336
Stage 1 representation width	384
Entity queries per view	48
Scene memory slots	16
Question-token length	256
ROI budget / crop size	4 / 224
Active evidence steps	6
Maximum views (model)	16
SCVR initial evidence budget	8
SCVR maximum evidence budget	24
Verify-expand increment	4
Training hardware / wall time	TBD
Final downstream reasoners	TBD