Representing Image-Required Mathematics for Intelligent Textbooks: A Controlled Comparison of Diagram Encodings

Ethan Croteau[1], Neil Heffernan[1]

[1]*Worcester Polytechnic Institute, Worcester, MA, USA*

Abstract

Digital mathematics textbooks increasingly support intelligent services such as tutoring, question answering, accessibility support, and content validation. Yet many mathematics problems depend on diagrams, graphs, number lines, geometric figures, or other visual content that is required for a correct solution. This paper studies how mathematics problems that require visual information can be represented for intelligent textbook systems. We conduct an exploratory controlled answer-scoring study with 34 middle-school mathematics problems, five diagram encodings, two model snapshots, and three attempts per problem-condition-model combination, yielding 1,020 scored runs. The encodings include original images, structured image-only descriptions that state visible diagram facts in text, reconstructed PNG images, SVG source supplied as text, and executable Python/Matplotlib reconstruction scripts. Results show that representation choice substantially affects performance. For GPT-5, accuracy rises from 55.9% with original images to 97.1% with structured text and 96.1% with executable reconstructions. GPT-5.5 has a stronger baseline with the original image (87.3%), but reaches 99–100% under structured text, SVG-as-text, and executable reconstruction conditions. An artifact audit shows that the highest-performing encodings often expose task-relevant structure more explicitly than learner-facing images. We therefore interpret the gains as evidence for both representation effects and human-authored semantic preprocessing, and argue that intelligent textbooks should treat diagram encodings as design choices whose appropriateness depends on the intended consumer and use case. We release code, prompt templates, artifacts, and analysis outputs for replication.

Supplementary materials: https://osf.io/5wbxd/

Keywords

intelligent textbooks, image-required mathematics, multimodal evaluation, diagram representation, accessibility, AI-ready educational content

1. Introduction

Digital textbooks increasingly support intelligent educational services rather than serving only as static containers of content. In mathematics, this shift creates a representational challenge: many textbook problems depend on visual content that is necessary for a correct solution. Diagrams, graphs, number lines, geometric figures, tables, and other visuals often encode information needed for measurement, counting, correspondence, spatial reasoning, or algebraic interpretation [1, 2, 3]. For intelligent textbook systems, the question is therefore not only whether a model can process an image, but what representation of that image the textbook should provide to support reliable, accessible, and auditable reasoning.

Rather than treating textbook figures only as raster inputs to a model, this work treats diagram encodings as textbook artifacts that can be authored, validated, and selected for intelligent services.

Prior work on image-required middle-school mathematics showed that multimodal models can vary substantially in their ability to solve authentic curriculum problems with and without images [4]. That finding motivates a broader intelligent-textbook question. If a textbook problem cannot be solved from text alone, should an AI service receive the original image, a structured textual description, a cleaned or reconstructed visual, a vector representation, or an executable description of the diagram? These encodings are not interchangeable [2]. Structured text may make counts and labels explicit;

iTextbooks'26: Seventh Workshop on Intelligent Textbooks, June 28, 2026, Seoul, Republic of Korea
ecroteau@wpi.edu (E. Croteau); nth@wpi.edu (N. Heffernan)

a reconstructed PNG may preserve the learner-facing visual form; SVG source may expose vector structure; and executable reconstructions may support authoring, validation, and re-rendering. Which encoding is appropriate depends on whether it is intended for a learner, an accessibility tool, an AI tutor, an evaluation pipeline, or an authoring workflow.

This paper studies these alternatives as diagram encodings for intelligent textbooks. We focus on mathematics problems that cannot be solved from text alone because they provide a concrete testbed for comparing diagram encodings. The empirical task is deliberately narrower than the full tutoring or accessibility setting: we evaluate answer-scored model problem solving, then discuss what the results imply and do not imply for downstream textbook services. We use *semantic enrichment* to refer to how explicitly a representation exposes task-relevant structure that a solver would otherwise need to extract from the visual diagram. We ask three questions:

1. How does model performance vary when textbook problems that require diagram information are presented using different diagram encodings?
2. How does semantic enrichment in non-image representations relate to model performance?
3. What do problem-level representation effects suggest for representing diagrams in intelligent textbooks?

The paper makes three contributions. First, it articulates a service-oriented representation framework that treats mathematics diagrams needed for solving as textbook artifacts to be authored, audited, stored, and selected for downstream intelligent services. Second, it reports a controlled representation-comparison study across 34 problems, five diagram encodings, two models, and 1,020 scored runs, with supplementary materials supporting replication. Third, it introduces semantic-enrichment, uncertainty, and problem-level effect analyses showing that alternative encodings can improve answer-scoring performance while also changing task access, motivating service-specific representation choices rather than a single best encoding.

2. Background and Motivation

2.1. Image-required mathematics in intelligent textbooks

Recent iTextbooks work positions digital textbooks not only as repositories of content, but also as supports for intelligent services such as help seeking, explanatory question answering, multimodal retrieval, accessibility support, and interactive knowledge navigation [5, 6, 7, 8, 9]. These services depend on the textbook content being available in forms that can be inspected, retrieved, adapted, and used by computational systems.

Mathematics introduces a particular challenge because some textbook visuals are not merely decorative or supportive, but required for a correct solution. A problem may require a learner or model to read a plotted value, count objects, measure from a scale, identify a corresponding side, interpret a hanger diagram, or infer a geometric relation from a figure. In such cases, omitting the image changes the task. Prior work on image-required middle-school mathematics showed that multimodal models vary substantially when authentic curriculum problems are evaluated with and without images [4]. More broadly, visual-math benchmarks such as MathVista, MathVerse, and We-Math show that diagram understanding and quantitative visual reasoning remain challenging for multimodal models [10, 11, 12]. Recent work on classroom-authentic visual reasoning and geometric shape perception points to similar limits in educational and spatial settings [13, 14]. A digital textbook that exposes only the problem text to an intelligent service therefore risks presenting an incomplete problem.

This distinction matters for textbook services that operate over visual content, including tutoring, question answering, accessibility, assessment, and authoring. Intelligent tutoring systems are a long-standing context for AI-supported learning [15]; in visual mathematics, a tutoring system may need to answer a student's question about a diagram. An accessibility tool may need to describe an image

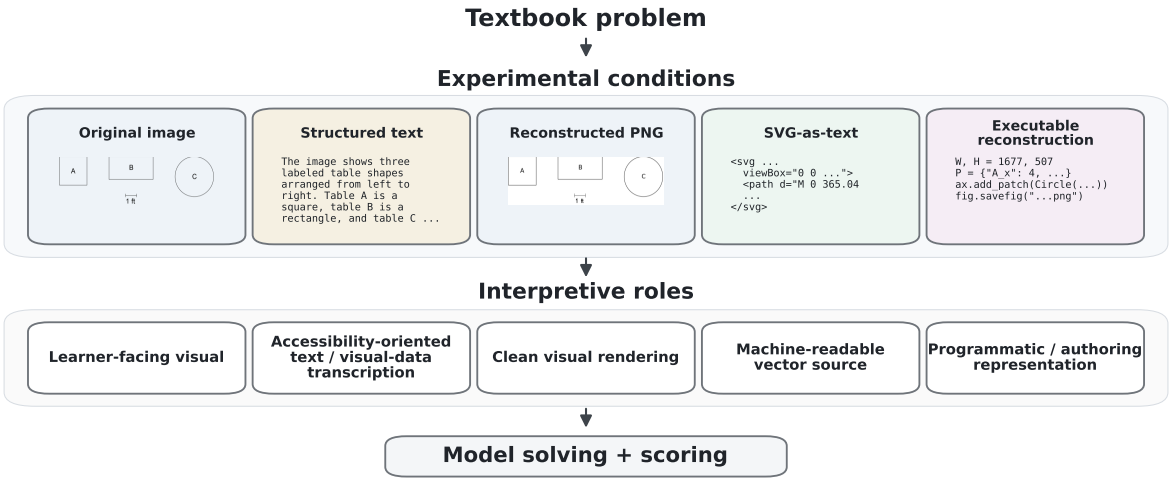


Figure 1: Representation framework used in the study. The five experimental conditions are illustrated using one example problem. The row labeled "Interpretive roles" summarizes possible uses for these encodings in intelligent textbooks; these roles are not intended as a one-to-one mapping to the experimental conditions.

without giving away the solution; an assessment system may need to verify that the expected answer is supported by the visible content; and a content-authoring system may need to regenerate or adapt a figure for different display contexts. These uses require more than detecting that an image exists: they require knowing what task-relevant information the visual content contributes.

2.2. From static images to machine-actionable representations

Prior iTextbooks work has examined how textbook content can be modeled, extracted, retrieved, and enriched for intelligent services. For example, textbook modeling and PDF extraction work has focused on recovering layout and activity structure from static textbook pages [16], while multimodal retrieval work has explored how digital textbook and classroom data can support LLM-based retrieval over audio, visual, and textual sources [7]. Grounded explanatory AI highlights the need for textbook services that respond using reliable source content rather than generated information that is unsupported or inaccurate [6]. Help-seeking systems similarly show how extracted textbook content can support context-specific recommendations within an iTextbook [5].

Our work focuses on a complementary representation problem: how mathematics diagrams that carry information needed for solving should be encoded for intelligent textbook systems. For the services considered here, a static image is useful for human viewing but limited as a machine-facing artifact: it is difficult to inspect, validate, adapt, search, or re-render without additional structure. Alternative encodings can expose different aspects of the same visual content. Structured image-text representations can externalize labels, counts, relations, and visible measurements. Reconstructed PNGs can provide a cleaner visual presentation while preserving a familiar learner-facing form. SVG source can represent visual elements in a text-readable vector format, a direction echoed by recent work treating SVG as a symbolic visual representation [17]. Executable reconstructions can support reproducible rendering and content validation; recent work on symbolic graphics programs similarly treats programmatic graphics as a substrate for assessing spatial-semantic reasoning in language models [18].

Figure 1 summarizes this framing. The central premise of this paper is that these representations have different affordances for intelligent textbooks. The goal is not to identify one universally best representation, but to understand how representation form and semantic explicitness affect model use of visual mathematics content. This framing treats diagram encodings as design choices for AI-supported textbook services rather than as interchangeable substitutes for learner-facing images.

3. Method

We conducted an exploratory controlled representation study over 34 mathematics problems requiring diagram information. The experiment holds the original problem text fixed while varying only the representation of the diagram. We compare five diagram encodings across two models, with three attempts for each problem-condition-model combination, yielding 1,020 scored runs.

3.1. Representation conditions

We compare five diagram encodings, summarized in Table 1. The conditions are designed as representation choices for intelligent textbook services rather than as prompt variants: each corresponds to a different way an intelligent textbook might expose diagram content to learners, accessibility tools, AI tutors, evaluation pipelines, or authoring systems. They are therefore not intended to be visually equivalent substitutes.

Table 1
Diagram encodings compared in the experiment.

Encoding	What the model receives	What the condition tests
Original image	Problem text plus the source textbook image	Baseline learner-facing visual representation.
Structured text	Problem text plus an image-focused structured description	Whether task-relevant visual information can be made explicit in text.
Reconstructed PNG	Problem text plus a rendered reconstruction image	Whether a faithful cleaned visual replacement supports reasoning.
SVG-as-text	Problem text plus SVG source for the reconstructed figure	Whether vector/text diagram structure is useful as a machine-readable representation.
Executable reconstruction	Problem text plus a neutral Python/Matplotlib reconstruction script	Whether programmatic diagram representations support reasoning and validation.

Original images and reconstructed PNGs are visual representations. Structured text provides the diagram information in written form. SVG-as-text and executable reconstruction are machine-actionable representations that may be more useful for AI services or authoring workflows than for direct learner presentation.

3.2. Artifact authoring and semantic-enrichment audit

For each selected problem, the source textbook text and original image define the reference task. We then created an artifact bundle containing problem metadata, the student-facing problem text, the original image, a structured image-text representation, an executable reconstruction script, and rendered PNG and SVG versions of the reconstruction. The structured image-text representation was authored to be image-focused and task-complete [19, 20, 21]: it describes the figure without restating the problem text, providing a solution strategy, or including the expected answer. The executable reconstruction was authored to be neutral and auditable: it recreates the diagram using explicit drawing primitives and avoids answer-bearing comments or variable names.

The artifact workflow proceeds from source problem to reusable artifact bundle, fidelity and neutrality checks, model runs, scoring, and semantic-enrichment analysis. Before running models, we checked whether artifacts were faithful, task-complete, neutral, free of unnecessary solution-adjacent information, attentive to visible measurements, careful not to silently resolve ambiguities in the source, and reusable as PNG/SVG renderings. These checks help distinguish artifacts that are close to learner-facing visual replacements from artifacts that intentionally expose more structure for accessibility, tutoring,

authoring, or validation. Although the experiment operationalizes representations as condition-specific model inputs, the conditions are intended as textbook artifact types that could be stored, validated, surfaced, or hidden depending on the consuming service.

Artifact authoring and review were conducted by the study team. The team had access to the source problem text, original image, and expected answer from the benchmark record, so answer-key access is a construct-validity risk rather than something hidden from the analysis. We addressed this risk procedurally, not by claiming independence: artifact files were reviewed for neutrality, the expected answer was not included in model prompts, and potential solution-adjacent structure was recorded in the semantic-enrichment audit. The resulting artifacts should therefore be interpreted as human-authored textbook-service representations, not as automatically generated equivalent renderings of the same visual task.

During artifact review, we found that non-image representations often expose more structure than the original learner-facing image. For example, the structured image-text representation may externalize measurements or counts, SVG source may expose coordinates and object groupings, and code may reveal drawing primitives or data structures. We therefore audited artifacts for the structured text, SVG-as-text, and executable reconstruction conditions for semantic enrichment, operationalized as the extent to which they externalize task-relevant structure that a solver would otherwise need to extract visually.

The audit labels are curated annotations recorded in the supplementary analysis outputs. These annotations were assigned by reviewing the artifact files rather than inferred automatically from model outcomes. We use these labels as interpretive annotations, not as independently established ground truth, quality rankings, or evidence of accessibility effectiveness. Low-enrichment artifacts primarily expose visible labels or rendering structure; medium-enrichment artifacts reduce visual extraction effort by making labels, coordinates, object ordering, dimensions, variables, or geometry easier to extract; high-enrichment artifacts externalize measured values, counts, inferred relations, groupings, comments, loops, or other machine-readable structure that can substantially simplify the task. The appropriateness of an enrichment level depends on the intended use: content shown directly to students may require less explicit structure, whereas accessibility tools, AI tutors, and authoring systems may benefit from more explicit representations.

3.3. Problem sample

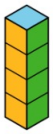
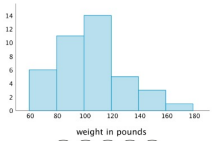
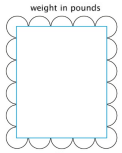
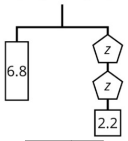
The source corpus is a 376-item image-required middle-school mathematics benchmark drawn from Illustrative Mathematics Grades 6–8 curriculum problems [22], delivered through ASSISTments [23]; benchmark construction is described in prior work [4]. We used an initial visual information demand audit to support stratified candidate selection, then manually selected and reviewed a 34-problem exploratory sample. We use *visual information demand* to describe the primary kind of visual work required to solve a problem, such as reading a scale, counting objects, extracting graph values, interpreting geometric structure, or using relational diagrams. Table 2 summarizes the six primary visual information demand categories used in this study, their counts in the selected sample, and one illustrative example per category.

The sample is purposeful rather than representative. It is designed to cover diverse representation challenges rather than mirror the full benchmark distribution. Accordingly, we interpret results as evidence about representation effects in a diverse exploratory sample, not as benchmark-wide prevalence estimates.

3.4. Models, prompting, and scoring

We evaluate two pinned model snapshots: `gpt-5-2025-08-07` and `gpt-5.5-2026-04-23`. We refer to these as GPT-5 and GPT-5.5 for readability. Both models were run through the OpenAI Responses API using the same prompt schema, `answer_object_v1`, and the same reasoning settings: `effort = high` and `summary = detailed`. Image detail was set to `high` for conditions that used image inputs. We did

Table 2
Primary visual information demand categories in the selected sample. Example problem thumbnails are illustrative rather than exhaustive.

Category	Visual requirement	n	Example
Measurement and scale	Read or compare quantities from a scale, ruler, or marked display	4	
Counting and enumeration	Count objects, outcomes, groups, or spatial units	5	
Graphs, plots, and coordinate systems	Extract values, intervals, coordinates, or relations from plotted displays	10	
Geometric structure and spatial relations	Interpret shape, area, perimeter, surface area, similarity, or spatial correspondence	8	
Algebraic/relational diagrams	Interpret equation-like visual structures such as bars, tapes, or hangers	4	
Tables and proportional representations	Use tabular or double-number-line structure to reason proportionally	3	

not explicitly configure temperature, top-p, random seed, or maximum output tokens, so those settings used the API defaults.

Each problem was evaluated under the five encodings in Table 1. For each condition, the original problem text was held constant and only the diagram representation changed. The same solver instruction and JSON response schema were used across all models, conditions, and attempts. The shared instruction asked the model to solve using the provided problem text and visual representation, state when information was missing, and return JSON containing solvability, required information, reasoning steps, and the final answer. The only condition-specific handling was representation delivery: for the original image and reconstructed PNG conditions, the corresponding image was attached to the API request; for the structured text, SVG-as-text, and executable reconstruction conditions, the full artifact contents were included in the prompt. SVG and Python files were included in fenced code blocks and were not truncated, minified, summarized, or otherwise transformed. All artifacts were finalized before model runs. Semantic-enrichment annotations were assigned by reviewing the artifacts and were not included in model prompts.

Each problem-condition-model combination is run three times, producing $34 \times 5 \times 2 \times 3 = 1{,}020$ scored runs. Responses are scored using an answer-extraction and scoring pipeline adapted from the source benchmark protocol [4]. The scorer extracts the final answer value from the JSON response and compares it with the expected answer using the study scoring method, which supports exact matching, numeric equivalence, symbolic equivalence, common unit cleanup, and approximate-word cleanup. The final analysis contains 1,020 scored responses, with no parse failures and no excluded manifest rows. For reproducibility, the supplementary materials include an anonymized problem manifest and artifact files, the run manifest, a scored response table, audit annotations, paper-level analysis outputs, and scripts documenting the prompt templates, request construction, and scoring policy.

3.5. Analysis

We analyze results at three levels. First, we compare accuracy by model and representation condition. Second, we compare accuracy by semantic-enrichment level, using the curated audit labels described above. Third, we summarize problem-level representation effects relative to the same model's performance with the original image.

Because attempts are nested within problems, we summarize uncertainty at the problem level rather than treating the 1,020 runs as independent. Using the problem-condition summary table, we bootstrap the 34 problem rows with replacement within each model-condition cell and recompute accuracy from the three-attempt correct counts. For each alternative condition, we also compute a paired problem-level delta against the original-image condition for the same problem and model. These intervals are descriptive rather than population-level claims because the sample is purposeful. To check sensitivity to the selected mix of visual demands, we also compute category-macro accuracy by first averaging within each primary visual information demand category and then averaging the six category means equally. We do not report F1 or MCC because the outcome is final-answer correctness for each response, not a classification task with interpretable true-negative cases. Accuracy, paired deltas, demand-category macro averages, and problem-level effect types are therefore the primary descriptive summaries.

For the problem-level analysis, each problem-model pair is assigned to one of five effect categories: all representations correct on all attempts; original image wrong on all attempts but at least one alternative produced a correct attempt; original image sometimes correct and an alternative improved performance; original image correct on all attempts but at least one alternative reduced performance; or all representations wrong. This analysis distinguishes broad condition-level gains from item-level patterns, including problems that are robust across representations, problems with correct attempts only after changing the representation, problems that improve under an alternative representation, and problems where changing the representation makes performance worse.

4. Results

4.1. Accuracy varies by diagram encoding

Table 3 shows that representation form substantially affects performance. For GPT-5, original images yield 55.9% accuracy, while structured text and executable reconstructions yield 97.1% and 96.1%, respectively. SVG-as-text also performs strongly at 91.2%. Reconstructed PNGs improve over the original image condition, but more modestly, reaching 63.7%. The paired problem-level bootstrap intervals show the same pattern: structured text, SVG-as-text, and executable reconstruction have large positive deltas against the original-image baseline, while the reconstructed-PNG delta is small and its interval includes zero. GPT-5.5 has a stronger baseline with the original image (87.3%), but still reaches 99–100% under structured text, SVG-as-text, and executable reconstruction conditions.

In this study, the strongest performance improvements do not come merely from replacing the original image with a cleaner visual rendering. Reconstructed PNGs help, but performance improves more when the representation exposes task-relevant structure in textual, vector, or programmatic form.

Because the selected problems span six primary visual information demand categories, we also examine whether representation effects are consistent across demand types.

Figure 2 shows that the representation effects vary across visual information demands as well as across models. Demand-category macro averages preserve the overall ordering despite the purposeful sample: for GPT-5 the macro accuracies are 60.8% for original image, 97.9% for structured text, 66.1% for reconstructed PNG, 91.1% for SVG-as-text, and 97.2% for executable reconstruction; for GPT-5.5 they are 87.2%, 99.3%, 93.3%, 100.0%, and 100.0%, respectively. These macro summaries should be read cautiously because some categories contain only three or four problems, but they reduce the risk that the headline pattern is only a consequence of over-weighting one visual-demand category.

Table 3
Accuracy by representation condition with problem-cluster bootstrap intervals. Each model-condition cell reports 102 runs, corresponding to 34 problems with three attempts each. CIs resample problem rows, not individual attempts. Delta is paired against the same model's original-image condition.

		Accuracy			Paired Δ vs. original		
Condition	Correct	%	CI low	CI high	pp	CI low	CI high
GPT-5							
Original image	57	55.9	40.2	71.6	–	–	–
Structured text	99	97.1	92.2	100.0	+41.2	+25.5	+56.9
Reconstructed PNG	65	63.7	48.0	78.4	+7.8	-2.0	+19.6
SVG-as-text	93	91.2	82.4	98.0	+35.3	+21.6	+50.0
Executable reconstruction	98	96.1	89.2	100.0	+40.2	+24.5	+55.9
GPT-5.5							
Original image	89	87.3	75.5	97.1	–	–	–
Structured text	101	99.0	97.1	100.0	+11.8	+2.9	+22.5
Reconstructed PNG	96	94.1	87.3	99.0	+6.9	-2.9	+17.6
SVG-as-text	102	100.0	100.0	100.0	+12.7	+2.9	+24.5
Executable reconstruction	102	100.0	100.0	100.0	+12.7	+2.9	+24.5

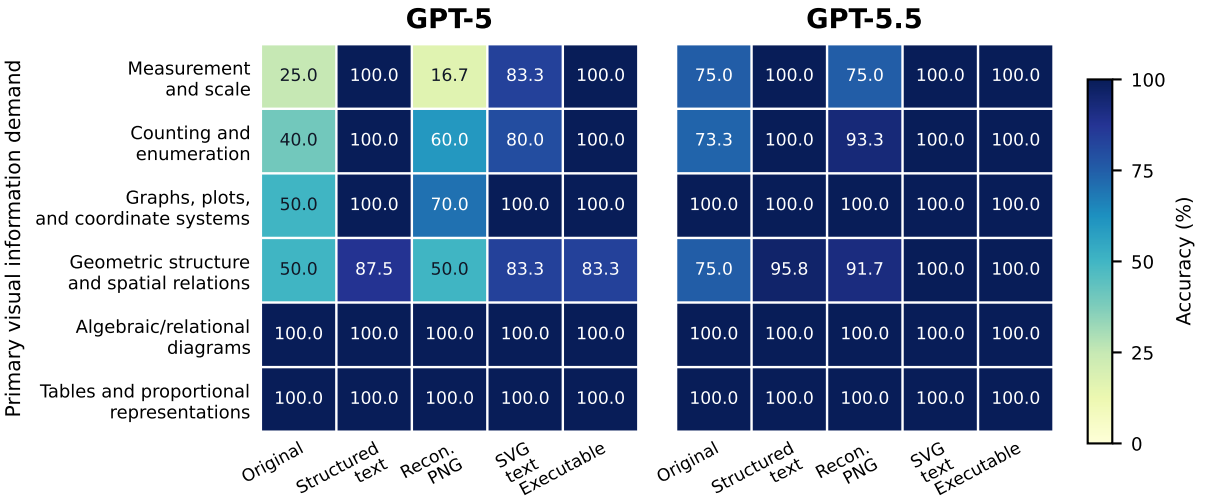


Figure 2: Accuracy by primary visual information demand and representation condition. Cells show percentage correct across all attempts for problems in each selected category.

4.2. Semantic enrichment is associated with much of the gain

Table 4 groups representation conditions by semantic-enrichment level. The baseline level corresponds to the original image condition. Reconstructed PNGs are treated as low-enrichment visual alternatives, while artifacts for the structured text, SVG-as-text, and executable reconstruction conditions receive low, medium, or high labels from the curated artifact audit.

Low-enrichment representations improve over the baseline visual condition for both models, but medium- and high-enrichment representations show much higher accuracy. The relationship is not strictly monotonic: for GPT-5, medium-enrichment artifacts outperform high-enrichment artifacts. This suggests that explicit structure is useful, but more semantic exposure is not automatically better and should be interpreted relative to the intended use of the representation.

This finding is important for intelligent textbooks because it separates visual clarity from semantic explicitness. A reconstructed image can improve the visual presentation, but semantically enriched encodings can change what information is directly available to the model. Those encodings may be appropriate for accessibility, tutoring, or machine-actionable services, but they should not be interpreted as visually equivalent substitutes for the original image or treated as universally better representations.

Table 4
Accuracy by semantic-enrichment level. Baseline visual corresponds to the original image condition; other levels come from the artifact audit. "Prob." is the number of problems represented at that level, and "Cond." is the number of problem-condition pairs.

				GPT-5		GPT-5.5	
Level	Prob.	Cond.	Runs	Correct	Acc. (%)	Correct	Acc. (%)
Baseline	34	34	102	57	55.9	89	87.3
Low	34	48	144	107	74.3	138	95.8
Medium	23	45	135	130	96.3	135	100.0
High	19	43	129	118	91.5	128	99.2

Table 5
Problem-level representation effect types. Counts are out of 34 problems per model.

Model	N	All correct	Alt. helped	Improved by alt.	Reduced by alt.	None correct
GPT-5	34	16	12	5	1	0
GPT-5.5	34	27	3	2	2	0

All correct indicates that every representation condition was correct on all three attempts. *Alt. helped* indicates that the original image condition was wrong on all three attempts and at least one alternative produced a correct attempt. *Improved by alt.* indicates that the original image condition was correct on one or two attempts and an alternative improved performance. *Reduced by alt.* indicates that the original image condition was correct on all three attempts but at least one alternative was not. *None correct* indicates that no representation condition produced a correct attempt.

4.3. Problem-level effects are heterogeneous

Table 5 summarizes problem-level representation effects. For GPT-5, 16 of 34 problems are correct across all representations, while 12 have at least one correct attempt from an alternative representation after failing with the original image and 5 improve from sometimes correct performance with the original image. For GPT-5.5, 27 of 34 problems are correct across all representations, but alternatives still help in 5 cases. No problem is wrong across all representations for either model. However, alternative representations can also reduce performance in otherwise correct cases, showing that encodings require validation rather than being treated as interchangeable substitutes.

This problem-level view is important because condition-level averages hide distinct representation effects. Some problems are robust across encodings, some produce correct attempts only when structure is externalized, and some are sensitive to the form of the alternative representation.

4.4. Representative cases illustrate affordances and risks

We highlight four representative cases to illustrate the range of problem-level effects (Table 6). The example in Figure 1 corresponds to Case 1, a measurement-and-scale item asking how many customers should be seated at table A, with expected answer 3. The structured description for this case states that table A has a side length of about 2.5 ft, which is visible in principle but is also much closer to the needed area calculation than the original learner-facing image. This is exactly the construct-validity issue raised by the semantic-enrichment audit: alternative encodings may improve accuracy by externalizing human-measured or human-organized information, not merely by changing file format.

Case 2 is a geometric composite-boundary problem involving a scalloped picture frame. It remains difficult for GPT-5 across encodings, while GPT-5.5 improves from 0/3 with the original image to 3/3 under SVG-as-text and executable reconstruction. Case 3 is a ruler/tick-mark measurement item where a reconstructed PNG hurts both models relative to their original-image baselines, showing that even low-enrichment visual reconstructions require validation. Case 4 is a counting item where the reconstructed PNG helps, but structured and executable encodings also expose object-count structure more directly. The full supplementary package includes the original problem text, diagrams, structured descriptions, SVG/code artifacts, expected answers, scored response records, and audit notes for these and all other cases.

Table 6
Case-study summary of representation effects and design lessons.

Problem	Visual information demand	Representation effect	Design lesson
Case 1	Measurement and scale	GPT-5 was partly correct with the original image; structured text and executable reconstruction reached 3/3. GPT-5.5 failed the original image but reached 3/3 under all alternatives.	Accessibility-oriented or programmatic artifacts can recover difficult scale information, but may externalize measured quantities.
Case 2	Geometric structure and spatial relations	GPT-5 remained difficult across encodings, with only structured text reaching 1/3. GPT-5.5 improved from 0/3 with the original image to 3/3 under SVG-as-text and executable reconstruction.	Geometry encodings can expose construction structure, but benefits depend on model capability and artifact explicitness.
Case 3	Measurement and scale	GPT-5 improved from 1/3 with the original image to 3/3 under structured text, SVG-as-text, and executable reconstruction. Reconstructed PNG hurt both models relative to their original-image baselines.	Clean visual reconstructions still require validation; low-enrichment alternatives are not guaranteed substitutes.
Case 4	Counting and enumeration	GPT-5 failed the original image but reached 3/3 under structured text, reconstructed PNG, and executable reconstruction. GPT-5.5 improved from 1/3 with the original image to 3/3 under all alternatives.	Counting tasks may benefit from visual cleanup, while text/code artifacts can expose counts or object structure more directly.

5. Discussion

The results support the central claim that diagram encodings are design choices for intelligent textbooks. Original images, reconstructed images, structured descriptions, SVG source, and executable reconstructions are not interchangeable: they expose different information, support different forms of machine access, and can affect model performance in different ways. The evidence is strongest for answer-scored problem solving, because that is what the experiment directly measures. Claims about tutoring, accessibility, authoring, and validation should therefore be read as design implications that require service-specific evaluation.

For intelligent textbooks, the practical implication is that diagram representations should be authored, validated, and selected according to the consuming service rather than assumed to be interchangeable views of the same content.

The contrast between reconstructed PNGs and semantically enriched representations is especially important. Reconstructed PNGs preserve a learner-facing visual form and provide modest improvements, whereas structured text, SVG-as-text, and executable reconstructions produce stronger performance improvements, especially for GPT-5. This suggests that intelligent textbook services may benefit from representations that expose task-relevant structure rather than relying only on raster image interpretation. At the same time, these enriched representations can change the nature of the task. A structured description that states counts or measured values may be appropriate for accessibility, tutoring support, or content validation, but it is not equivalent to asking a learner or model to extract the same information visually.

The semantic-enrichment audit therefore shapes how the results should be interpreted. Higher performance under structured, vector, or executable encodings should not be read simply as evidence that one representation is better. Instead, these encodings provide different affordances for different consumers. Some are closer to learner-facing alternatives, while others are better understood as AI-facing metadata or authoring and validation artifacts. Intelligent textbook systems may therefore need layered representations: a visual diagram for learners, a structured textual description for accessibility and question answering, a vector representation for machine inspection, and executable source for authoring or validation.

These findings suggest that diagram encodings should be treated as service-specific design choices. Learner-facing content may prioritize visual fidelity and limited semantic exposure, while accessibility support, diagram-aware tutoring, and question answering may benefit from more explicit structured

descriptions. However, a tutoring system needs more than correct final answers: it must ground hints in diagram parts, respond to student reasoning, manage multi-turn uncertainty, and avoid over-disclosing solution steps. Similarly, an accessibility representation should be evaluated with accessibility expertise and users rather than inferred from model accuracy. Authoring and content-validation workflows may benefit from vector or executable representations that expose geometry, labels, grouping, and rendering structure. These uses are not mutually exclusive, but they show why representation choice should not be treated as a single-format optimization problem.

The model comparison also matters. GPT-5 shows strong representation sensitivity, while GPT-5.5 is more robust across encodings. This suggests that representation design may be especially important for models with imperfect visual extraction, but remains relevant even for stronger systems. More broadly, the results indicate that AI-ready textbook content should not be evaluated only against a single model snapshot. As models change, the role of representation may shift from enabling correctness to improving reliability, auditability, accessibility, and maintainability.

6. Limitations and Future Work

This study is exploratory. The 34-problem sample was purposefully selected to span visual information demands and should not be interpreted as representative of the full 376-item benchmark. Problem-cluster bootstrap intervals and category-macro averages make the descriptive uncertainty clearer, but the sample is still too small for strong population claims. Mixed-effects logistic regression is a natural future analysis for a larger corpus; in this pilot, near-ceiling cells and sparse demand categories make such modeling easy to over-interpret.

The central construct-validity limitation is that several conditions differ in both encoding format and amount of human-authored semantic preprocessing. Structured text, SVG source, and executable code sometimes expose counts, measurements, coordinates, construction logic, or object groupings that the original image requires a solver to infer visually. This is not simply a flaw: such explicit structure may be exactly what an accessibility, authoring, or validation service needs. It does mean that accuracy gains should be interpreted as effects of representation plus semantic enrichment, not as proof that one file format is a better equivalent rendering of the same task.

The semantic-enrichment labels are curated annotations, not independent ground-truth measurements or quality rankings. We did not collect multiple independent annotations or inter-rater agreement for the pilot labels. The artifact workflow was human-guided and conducted by the study team, which had access to expected answers during authoring and review. Future work should separate authoring and auditing roles more strictly or use blind secondary review to assess leakage and solution-adjacent structure.

The study evaluates answer-scored model problem solving as a proxy for representation usefulness. Improved model performance does not by itself establish learner benefit, accessibility quality, tutoring effectiveness, or deployment readiness. The model comparison is also limited to two pinned GPT snapshots; additional studies should test other proprietary, open-weight, and vision-specialized systems before treating the observed representation effects as model-general.

Future work should expand the sample, validate visual information demand and semantic-enrichment labels with multiple reviewers, and compare learner-facing, accessibility-oriented, AI-facing, and authoring-oriented variants of the same artifacts. Such work would help separate visual cleanup, explicit data transcription, vector structure, and programmatic representation. It should also evaluate diagram-aware question answering, tutoring, accessible rendering, authoring support, related-problem generation, and content-quality auditing directly. Dynamic visual-math benchmarks similarly motivate robustness evaluation across related visual and textual variants [24]. Future studies should test not only answer accuracy, but also grounded hints, explanations, scaffolds, refusal behavior, and whether remaining failures reflect model limitations, artifact defects, scoring assumptions, or source-item quality issues.

7. Conclusion

Intelligent textbooks need representations of visual mathematics content that go beyond static images alone. In this exploratory controlled study, diagram encodings substantially affected answer-scored model performance across 34 mathematics problems that require visual information. Structured text, SVG-as-text, and executable reconstructions often improved performance more than reconstructed PNGs, but these gains were associated partly with human-authored semantic enrichment rather than visual equivalence. The results suggest that intelligent textbooks should expose multiple, validated representations of diagram content and treat those representations as design choices for AI-supported learning services, not as interchangeable substitutes or universal replacements for learner-facing images.

8. Acknowledgments

We would like to thank NSF (e.g., 2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (e.g., R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, & R305A120125), GAANN (e.g., P200A180088 & P200A150306), EIR (U411B190024 & S411B210024), ONR (N00014-18-1-2768), NIH (R44GM146483), and Schmidt Futures. None of the opinions expressed here are those of the funders.

Declaration on Generative AI

During preparation, the authors used OpenAI ChatGPT and Codex for grammar/spelling checks, paraphrase/rewording, writing-style suggestions, LaTeX table formatting, and reviewer-response/submission checklist review. The authors reviewed and edited all AI-assisted content, verified the analyses and claims, and take full responsibility for the publication's content.

References

[1] J. H. Larkin, H. A. Simon, Why a diagram is (sometimes) worth ten thousand words, Cognitive Science 11 (1987) 65–100.

[2] S. Ainsworth, DeFT: A conceptual framework for considering learning with multiple representations, Learning and Instruction 16 (2006) 183–198.

[3] R. E. Mayer, Multimedia Learning, 3rd ed., Cambridge University Press, Cambridge, UK, 2020.

[4] E. Croteau, N. Heffernan, Seeing is solving: MLLMs, reasoning, and refusal in visual math, Journal of Educational Data Mining 18 (2026) 244–285. doi:10.5281/zenodo.19420820.

[5] Y.-J. Tseng, Y.-H. Lin, G. Yadav, N. Bier, V. Aleven, Curio: An on-demand help-seeking system on itextbooks for accelerating research on educational recommendation algorithms, in: Proceedings of the Fifth International Workshop on Intelligent Textbooks 2023 co-located with the 24th International Conference on Artificial Intelligence in Education (AIED 2023), volume 3444 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 94–101. URL: https://ceur-ws.org/Vol-3444/itb23_s4p3.pdf.

[6] F. Sovrano, K. Ashley, A. Bacchelli, Toward eliminating hallucinations: Gpt-based explanatory ai for intelligent textbooks and documentation, in: Proceedings of the Fifth International Workshop on Intelligent Textbooks 2023 co-located with the 24th International Conference on Artificial Intelligence in Education (AIED 2023), volume 3444 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 54–65. URL: https://ceur-ws.org/Vol-3444/itb23_s3p2.pdf.

[7] B. Wright, V. Guruvayur, L. Napolitano, D. Ozar, A. Rivera, A. Sai, B. Tafesse, Using digital textbook and classroom data to explore multimodal (audio, visual, & textual) llm retrieval techniques, in: Proceedings of the Sixth International Workshop on Intelligent Textbooks 2025 co-located with the 26th International Conference on Artificial Intelligence in Education (AIED 2025), volume

4010 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 59–70. URL: https://ceur-ws.org/Vol-4010/itb25_s2s3.pdf.

[8] B. Johnson, B. Jerome, J. Dittel, R. V. Campenhout, Improving textbook accessibility through ai simplification: Readability improvements and meaning preservation, in: Proceedings of the Sixth International Workshop on Intelligent Textbooks 2025 co-located with the 26th International Conference on Artificial Intelligence in Education (AIED 2025), volume 4010 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 45–58. URL: https://ceur-ws.org/Vol-4010/itb25_s2p2.pdf.

[9] S. Tytenko, Ai-driven interactive hierarchical concept maps for digital learning environments and intelligent textbooks, in: Proceedings of the Sixth International Workshop on Intelligent Textbooks 2025 co-located with the 26th International Conference on Artificial Intelligence in Education (AIED 2025), volume 4010 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 3–16. URL: https://ceur-ws.org/Vol-4010/itb25_s1p1.pdf.

[10] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, J. Gao, MathVista: Evaluating mathematical reasoning of foundation models in visual contexts, The Twelfth International Conference on Learning Representations (ICLR 2024), OpenReview, 2024. URL: https://openreview.net/forum?id=KUNzEQMWU7.

[11] R. Zhang, D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K.-W. Chang, Y. Qiao, et al., MATHVERSE: Does your multi-modal LLM truly see the diagrams in visual math problems?, in: European Conference on Computer Vision, Springer Nature Switzerland, Cham, 2024, pp. 169–186.

[12] R. Qiao, Q. Tan, G. Dong, M. MinhuiWu, C. Sun, X. Song, J. Wang, Z. GongQue, S. Lei, Y. Zhang, Z. Wei, M. Zhang, R. Qiao, X. Zong, Y. Xu, P. Yang, Z. Bao, M. Diao, C. Li, H. Zhang, We-Math: Does your large multimodal model achieve human-like mathematical reasoning?, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 20023–20070. URL: https://aclanthology.org/2025.acl-long.983/. doi:10.18653/v1/2025.acl-long.983.

[13] M. Huti, A. Mackintosh, A. Waldock, D. Andrews, M. Lelièvre, M. Boos, T. Murray, P. Atherton, R. A. A. Ince, O. G. B. Garrod, Visual Reasoning Benchmark: Evaluating multimodal LLMs on classroom-authentic visual problems from primary education, 2026. URL: https://arxiv.org/abs/2602.12196. doi:10.48550/arXiv.2602.12196. arXiv:2602.12196.

[14] W. Rudman, M. Golovanevsky, A. Bar, V. Palit, Y. LeCun, C. Eickhoff, R. Singh, Forgotten Polygons: Multimodal large language models are shape-blind, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 11983–11998. URL: https://aclanthology.org/2025.findings-acl.620/. doi:10.18653/v1/2025.findings-acl.620.

[15] K. VanLehn, The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems, Educational Psychologist 46 (2011) 197–221.

[16] É. Lincker, O. Pons, C. Guinaudeau, I. Barbet, J. Dupire, C. Hudelot, V. Mousseau, C. Huron, Layout- and activity-based textbook modeling for automatic pdf textbook extraction, in: Proceedings of the Fifth International Workshop on Intelligent Textbooks 2023 co-located with the 24th International Conference on Artificial Intelligence in Education (AIED 2023), volume 3444 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 37–53. URL: https://ceur-ws.org/Vol-3444/itb23_s3p1.pdf.

[17] K. Q. Lin, Y. Zheng, H. Ran, D. Zhu, D. Mao, L. Li, P. Torr, A. J. Wang, VCode: a multimodal coding benchmark with SVG as symbolic visual representation, 2025. URL: https://arxiv.org/abs/2511.02778. doi:10.48550/arXiv.2511.02778. arXiv:2511.02778.

[18] Z. Qiu, W. Liu, H. Feng, Z. Liu, T. Xiao, K. Collins, J. B. Tenenbaum, A. Weller, M. J. Black, B. Schölkopf, Can large language models understand symbolic graphics programs?, in: Y. Yue, A. Garg, N. Peng, F. Sha, R. Yu (Eds.), International Conference on Learning Representations, volume 2025, 2025, pp. 26265–26311. URL: https://proceedings.iclr.cc/paper_files/paper/2025/file/41bd71e7bf7f9fe68f1c936940fd06bd-Paper-Conference.pdf.

[19] W3C Web Accessibility Initiative, Images tutorial: Complex images, https://www.w3.org/WAI/tutorials/images/complex/, 2022. Updated 17 January 2022.

[20] DIAGRAM Center, Image description guidelines, https://diagramcenter.org/making-images-accessible.html, 2015.

[21] C. Yan, H.-P. Hutter, F. M. Schmitt-Koopmann, A. Darvishy, Chart Accessibility: A review of current alt text generation, IEEE Access 13 (2025) 94040–94056. doi:10.1109/ACCESS.2025.3571626.

[22] I. Mathematics, Illustrative Mathematics, grade 6–8, Available at https://illustrativemathematics.org/, 2019. Authored by Illustrative Mathematics.

[23] N. T. Heffernan, C. L. Heffernan, The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching, International Journal of Artificial Intelligence in Education 24 (2014) 470–497. doi:10.1007/s40593-014-0024-x.

[24] C. Zou, X. Guo, R. Yang, J. Zhang, B. Hu, H. Zhang, DynaMath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models, in: Y. Yue, A. Garg, N. Peng, F. Sha, R. Yu (Eds.), International Conference on Learning Representations, volume 2025, 2025, pp. 48337–48383. URL: https://proceedings.iclr.cc/paper_files/paper/2025/file/78b248ea6f627431bba5029d92be8a3d-Paper-Conference.pdf.