Trace: A Taxonomy-Guided Environment for Multidomain Visual Reasoning

Md Tanvirul Alam
Rochester Institute of Technology
Rochester, NY, USA
ma8235@rit.edu
https://maveryn.github.io/trace/

Abstract

Reinforcement learning with verifiable rewards (RLVR) has substantially improved language-model reasoning, yet its extension to vision-language models remains constrained by the lack of training data that are simultaneously broad, exactly verifiable, and reproducible. We introduce TRACE, a taxonomy-guided environment for multidomain visual reasoning. TRACE factorizes task construction into a scene grammar and an executable task program, separating visual realization from answer computation. A shared semantic state determines the rendered image, prompt, typed answer, verifier state, and replayable instance trace. The resulting environment comprises 1,000 tasks over 277 scene grammars and 11 visual domains, with controlled semantic and visual variation. RLVR on 64,000 TRACE instances improves the macro-average across 24 external benchmarks by 3.51 percentage points for Qwen2.5-VL-3B and 4.06 points for Qwen2.5-VL-7B, providing evidence that broad procedural training can transfer beyond the generated task distributions. **Project page:** https://maveryn.github.io/trace/

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has emerged as an effective approach for improving language-model reasoning, particularly in mathematical and programmatic domains where solutions can be evaluated with inexpensive, objective feedback [13, 30]. Extending this paradigm to vision-language models requires training data whose questions are grounded in visual content, whose answers can be verified exactly, and whose variation extends beyond a narrow set of templates. Recent multimodal RL methods report gains in visual mathematics, counting, spatial reasoning, and general visual understanding [15, 22, 25, 27], yet constructing visual training data that support broad and transferable reasoning remains a central bottleneck.

Existing work has pursued two main strategies. Large training mixtures broaden coverage by combining many datasets with task-specific rewards [29], but they inherit heterogeneous task granularity, sampling units, and reasoning boundaries from their source collections. Procedural systems instead provide exact supervision and controlled generation, but typically focus on a single scene grammar or a small set of closely related objectives, such as jigsaws, visual logic puzzles, games, or abstract perception tasks [1, 8, 38, 45, 51]. As a result, breadth, controllability, and exact verification are rarely combined within one coherent task organization.

We introduce TRACE, a taxonomy-guided environment that addresses this gap by separating visual construction from reasoning computation. A scene grammar defines the objects, relations, and layout family that can be rendered; a task program specifies the computation and answer type; and a reward contract defines exact verification. Bounded query variation changes a meaningful task argument, such as extremum direction, without creating a separate task. Building on structured

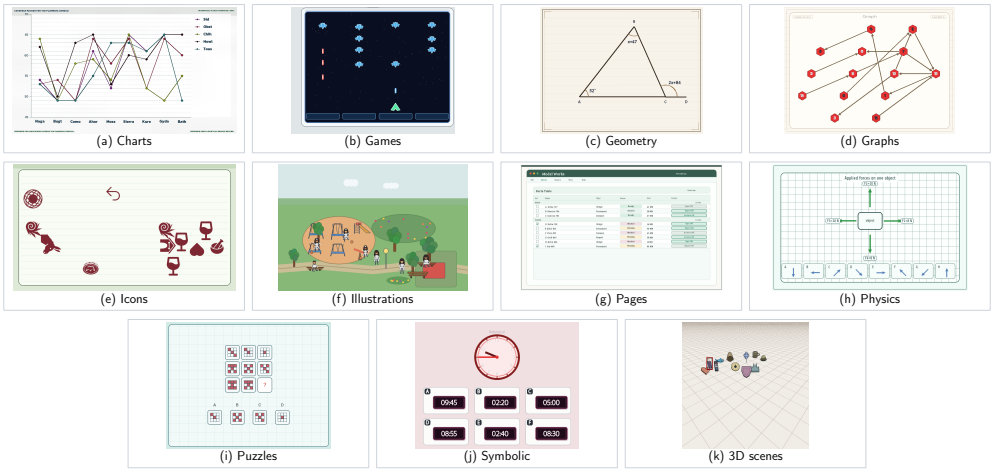


Figure 1: Representative instances from the 11 visual domains in TRACE, spanning charts, games, geometry, graphs, icons, illustrations, pages, physics, puzzles, symbolic notation, and three-dimensional scenes.

scene representations and executable question programs in CLEVR and Task Me Anything [18, 57], TRACE uses this decomposition not only to generate instances, but also to define the units used for sampling, verification, and cross-domain analysis.

This organization yields 1,000 tasks over 277 scene grammars spanning 11 visual domains. Each instance is generated from a semantic state on which the task program is executed; the same state determines the image, prompt, typed answer, and verifier state. Supervision therefore follows directly from the task computation rather than from labels assigned after rendering. The environment supports controlled variation in both semantic content and visual realization while preserving stable task identity and exact rewards. Figure 1 illustrates the resulting visual breadth.

We use TRACE to test whether RLVR on procedural visual tasks transfers beyond held-out instances from the same environment. Qwen2.5-VL-3B and Qwen2.5-VL-7B are trained on the same 64,000 TRACE instances and evaluated on 24 external visual-reasoning benchmarks. Training improves the benchmark macro-average by 3.51 percentage points at 3B and 4.06 points at 7B, providing evidence that broad procedural training can support transfer beyond the generated task distributions.

Our contributions are:

- A program-centered taxonomy that separates scene grammars, executable task programs, and bounded query variation, providing stable task units across 1,000 tasks, 277 scene grammars, and 11 visual domains;
- An executable generator–renderer–verifier environment in which images, prompts, typed answers, verifier states, and replayable instance traces derive from shared semantic state, enabling exact supervision and controlled semantic and visual variation; and
- A two-scale RLVR study showing gains of 3.51 and 4.06 points on the 24-benchmark macro-average at 3B and 7B, respectively, with positive mean changes on 21 of 24 benchmarks at 3B and all 24 benchmarks at 7B.

2 Related Work

Controlled visual task generation. Structured representations have long been used to make visual-reasoning datasets controllable and diagnostically useful. CLEVR generates synthetic scenes from explicit scene graphs and associates questions with functional programs, enabling controlled evaluation of compositional reasoning; GQA extends this paradigm to compositional questions grounded in real-image scene graphs [16, 18]. PuzzleVQA automatically constructs abstract-pattern questions with reasoning explanations and uses them to distinguish failures in visual perception,

inductive reasoning, and deductive reasoning [5]. Task Me Anything generalizes programmatic benchmark construction into an on-demand engine built around an extensible taxonomy of visual assets, attributes, relations, and task plans [57]. WorldBench instead uses a broad taxonomy of visual concepts to curate diverse real images and manually construct challenging questions that require reasoning beyond object recognition [53]. These works establish explicit structure as a basis for controlling evaluation coverage. TRACE builds on this lineage but makes the executable task definition a stable unit not only of question generation, but also of sampling, verification, and RL training.

Executable generator–verifier environments. Procedural reasoning environments replace fixed example collections with programs that generate instances and evaluate their solutions. Reasoning Gym provides more than one hundred generator–verifier environments across arithmetic, algebra, logic, graphs, geometry, computation, and games, with instance difficulty controlled through generation parameters [34]. Enigmata develops the same model for puzzle reasoning through 36 tasks in seven categories, each equipped with a controllable generator and rule-based verifier, and studies their use in multi-task RLVR [3]. These systems demonstrate how inspectable generation and exact rewards can support effectively unbounded training distributions. They primarily operate over textual or symbolic inputs, whereas TRACE extends the generator–verifier model to rendered visual states and separates the visual scene grammar from the task program executed over that state.

Multimodal RL data and mixture design. A major line of multimodal RL research constructs broad training mixtures from existing datasets and assigns rewards according to their answer formats. Visual-RFT, Vision-R1, LMM-R1, MM-Eureka, R1-Onevision, Reason-RFT, and VLM-R1 develop RL data and reward functions spanning classification, visual mathematics, counting, structural perception, spatial reasoning, and referring expressions [15, 22, 25, 27, 31, 37, 52]. Vero scales this approach to 600K examples drawn from 59 datasets across six task categories, using task-routed rewards to accommodate heterogeneous output spaces; its mixture ablations further indicate that broad category coverage is important for transfer [29]. More generally, OpenThoughts emphasizes that data sources, filtering, teacher generation, and mixture proportions should be treated as explicit and reproducible parts of a reasoning-data recipe [12].

A complementary line asks which examples and source proportions are most useful. SoTA with Less estimates sample difficulty through Monte Carlo tree search and retains a compact set of challenging examples, while RAP identifies high-value multimodal examples using discrepancies between multimodal and text-only behavior together with attention-based confidence [20, 44]. MoDoMoDo formulates multidomain RLVR as a mixture-optimization problem and learns to predict downstream training outcomes from the source distribution [21]. Vision-G1 combines data from 46 sources with influence-based selection, difficulty filtering, and curriculum training, whereas WeThink constructs a 120K-example mixture from 18 sources and combines rule-based and model-based rewards [50, 56]. DeepVision-103K broadens verified training data within visual mathematics through large-scale synthesis and curation [35]. Across these approaches, source selection, mixture weighting, and reward routing are central design variables. TRACE instead defines its sampling units through authored task programs before instances are generated, rather than inheriting task granularity from source datasets.

Synthetic and verifiable visual RL data. The most closely related systems create new visual experience specifically for reinforcement learning. Jigsaw-R1 uses jigsaw puzzles as a controlled testbed for studying visual RL, including generalization across puzzle configurations and the interaction between RL, supervised warm starts, and explicit reasoning [45]. VisualSphinx expands a small collection of puzzle rules through a rule-to-image pipeline that generates executable rendering code, grounded questions, and verifiable answers [8]. Game-RL adapts executable game code into visual states and verifiable tasks, thereby separating game environments from the objectives defined over them [38]. COGS takes a compositional approach: it decomposes seed questions into primitive perception and reasoning factors, recombines those factors with new images, and derives process-level rewards from the resulting intermediate subquestions and answers [11].

Other systems construct verifiable objectives by modifying existing data. SynthRL produces harder, answer-preserving variants of visual-mathematics questions and applies an explicit verification stage, whereas ViCrit injects a localized error into an otherwise valid image description and rewards the model for identifying the corrupted span [43, 47]. Sphinx procedurally generates 25 task families over motifs, tiles, charts, icons, and geometric primitives, with deterministic answers for both evaluation

and RLVR training [1]. Concurrent TRON provides 520 online generator–verifier environments across five ability groups, generating fresh instances during training and auditing generation reliability, diversity, and difficulty [51].

Game-RL distinguishes games from their associated tasks within one visual domain, while TRON treats each generator–verifier program as an online environment. TRACE extends the structural separation across 11 visual domains: 277 scene grammars define the states that can be rendered, while 1,000 stable task programs define the computations and reward contracts applied to those states. This factorization allows multiple reasoning objectives to share a scene grammar and allows visual realization to change without redefining the task being sampled, verified, or analyzed.

3 A Program-Centered Task Taxonomy

A procedural environment requires a stable unit of generation, sampling, verification, and analysis. Defining this unit is nontrivial: changes to the reasoning objective should produce distinct tasks, whereas changes to prompt arguments or visual realization should not artificially inflate the task inventory. TRACE addresses this problem through the hierarchy

$$\text{domain} \longrightarrow \text{scene grammar} \longrightarrow \text{task}. \tag{1}$$

A domain groups related visual inputs for reporting and mixture balancing. A scene grammar defines a reusable family of semantic states and visual realizations through its object vocabulary, relations, layout family, and question-facing scaffold. A task couples that grammar to an executable task program, answer schema, and reward contract. This factorization separates the visual structure of an instance from the computation used to answer it.

3.1 Task identity and equivalence

We represent a task as

$$T = (\mathcal{S}, P, \mathcal{Y}, \mathcal{C}), \tag{2}$$

where $\mathcal{S}$ is the scene grammar, P is the task program, $\mathcal{Y}$ is the answer schema, and $\mathcal{C}$ is the reward contract. The task program specifies how candidate objects are constructed, the roles of its operands, any intermediate computations, the final operator, and the binding between the computed result and the returned answer. Broad descriptions such as *count objects* or *select an extremum* therefore do not identify a task precisely enough for sampling or analysis.

Within TRACE, two candidate task definitions are merged only when

$$T_i \equiv T_j \quad \Longleftrightarrow \quad \mathcal{S}_i \simeq \mathcal{S}_j,\ P_i \simeq P_j,\ \mathcal{Y}_i = \mathcal{Y}_j,\ \mathcal{C}_i \simeq \mathcal{C}_j. \tag{3}$$

Here, $\mathcal{S}_i \simeq \mathcal{S}_j$ requires the same object vocabulary, relations, layout family, and visible scaffold; $P_i \simeq P_j$ requires the same candidate construction, operand roles, intermediate operations, final operator, and output binding; and $\mathcal{C}_i \simeq \mathcal{C}_j$ requires equivalent answer normalization and scoring behavior. These relations permit the renaming of literal operands, such as a color or category, but not changes to the underlying visual or computational structure. A change to any of these elements is represented as a separate task in TRACE, even when the resulting prompts appear similar or share the same answer type.

Table 1 summarizes the operational consequences of this rule.

3.2 Task, query, and generation boundaries

A task may expose an internal query variable q that selects a bounded argument of P while preserving the task program, answer schema, and reward contract. For example, choosing between the largest and smallest interquartile range changes the extremum direction but not the candidate set, metric, output binding, or verification rule.

Generation parameters θ instead determine the instance sampled from a fixed task definition. Semantic parameters vary the underlying state—such as object counts, values, thresholds, or board dimensions—and may therefore change the answer. Render-only parameters vary the visual realization of a fixed semantic state and query while preserving all answer-relevant relations and the resulting reward. Figure 2 illustrates these distinctions.

Table 1: Operational examples of the TRACE task-boundary rule. Changes to bounded program arguments remain queries, while semantic and render-only variation preserve task identity. Changes to the computation or scene grammar define a new task.

Variation	Effect on task definition	Classification
Palette or layout	Fixed semantic state, task program, and reward contract unchanged	Render-only generation
Item count or sampled values	Semantic state changes within the same scene grammar and task program	Semantic generation
Largest vs. smallest IQR	Only extremum direction changes	Query
Shape vs. shape-and-color count	Predicate arity and selected set change	New task
AND vs. inclusive OR count	Intermediate set operator changes	New task
Direct vs. algebraic exterior angle	Derivation rule changes	New task
Same extremum family in another scene grammar	Scene grammar changes	New task

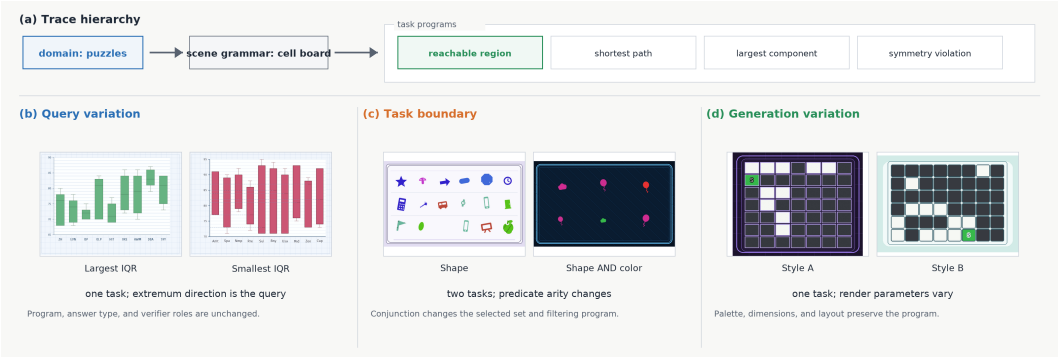


Figure 2: Examples of the task boundary in TRACE. **Left:** Changing extremum direction modifies only a bounded query argument and therefore remains within one boxplot task. **Center:** Adding a conjunctive attribute changes the candidate predicate and selected set, producing a new icon-filtering task. **Right:** Changes to board dimensions, layout, and visual treatment remain generation parameters because the reachability program and reward contract are unchanged.

This separation prevents both forms of inventory distortion. Treating every operand choice as a new task would overcount a single computation, whereas merging distinct intermediate operations would obscure meaningful differences in reasoning.

3.3 Canonical task programs and extensibility

Although task identity remains bound to a scene grammar, task-program operators are canonicalized across grammars so that recurring computational structures can be analyzed jointly. For example, selecting the boxplot with the extreme IQR and selecting the grid line with the extreme shape count both instantiate the canonical signature

$$\textit{filter candidates} \rightarrow \textit{compute metric} \rightarrow \textit{select extremum} \rightarrow \textit{return candidate label}.$$

They nevertheless remain distinct tasks because they operate over different scene grammars, construct different metrics, and bind the result to different visual entities. Canonicalization therefore exposes shared reasoning structure without collapsing scene-specific task identity.

The same distinction determines the training mixture. TRACE samples tasks uniformly by default and samples query values only after a task has been selected. Splitting a task into operand-specific variants would consequently overweight one program, whereas merging programs with different computations would hide heterogeneous reasoning behind a single sampling unit and aggregate score.

The factorization also provides a controlled path for extension. A new scene grammar introduces a new family of semantic states and visual realizations, while a new task adds an executable program and reward contract to an existing grammar. Renderers can therefore evolve without redefining the

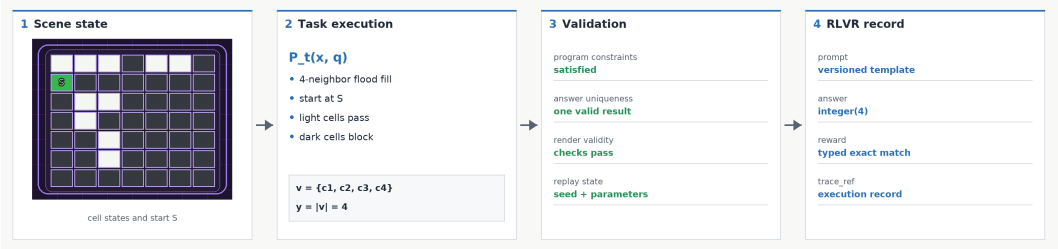


Figure 3: End-to-end construction of a TRACE instance. Four-neighbor reachability is computed over the semantic cell board, and the resulting state determines the typed answer, rendered image, prompt, scorer, and replayable instance trace.

tasks applied to their semantic states, and additional reasoning objectives can reuse an existing scene grammar without duplicating its visual construction.

4 The TRACE Environment

The taxonomy defines what constitutes a task; the environment instantiates that definition as a deterministic generator–renderer–verifier pipeline. For each sampled task, TRACE constructs a semantic state, executes the corresponding task program, renders the image and prompt, and binds the resulting answer and verifier state to an exact scorer. This design keeps generation, supervision, and replay grounded in the same underlying state.

4.1 Executable instance generation

Each generated instance is represented as

$$\mathcal{D}_i = \big(I_i, p_i, y_i, c_i, \tau_i\big), \tag{4}$$

where I_i is the rendered image, p_i is the prompt, y_i is the typed answer, c_i is the instance-bound scorer, and τ_i is the corresponding instance trace. The trace records the semantic state, sampled query, task execution, rendering decisions, and verifier state, making each instance replayable and supporting failure analysis.

Let s denote a scene grammar, t a task, q an internal query, z an instance seed, and θ the sampled generation parameters. Instance construction proceeds as

$$x = G_s(z, \theta_s), \qquad (y, v) = P_t(x; q), \tag{5}$$

$$I = R_s(x; z, \theta_r), \qquad p = H_{s,t}(x, q; z, \theta_p), \tag{6}$$

$$c = \mathrm{bind}(\mathcal{C}_t; y, v)\,, \qquad \tau = \Gamma(s, t, q, z, \theta, x, v). \tag{7}$$

The scene generator G_s first constructs semantic state x. The task program P_t then executes over that state and returns both the typed answer y and verifier state v. The renderer R_s and prompt function $H_{s,t}$ independently realize the image and question from the same semantic state. Finally, the task-level reward contract $\mathcal{C}_t$ is bound to the expected answer and verifier state to produce the instance scorer c, while Γ records the complete trace.

Semantic, prompt, and rendering choices are governed by independent seeded streams. The resulting instance can therefore be reconstructed from its recorded state without coupling superficial prompt or rendering variation to the task computation.

TRACE supports four answer interfaces: integers, canonical numeric values, strings, and option letters. Each reward contract specifies the corresponding type-aware normalization and comparison rule, which the bound scorer applies to the submitted answer.

Figure 3 illustrates the full pipeline for a cell-board task. A four-neighbor flood fill from the marked start cell produces the set of reachable passable cells, and the cardinality of that set gives $y = 4$. Passable cells outside the connected region remain visual distractors but do not enter the task computation.

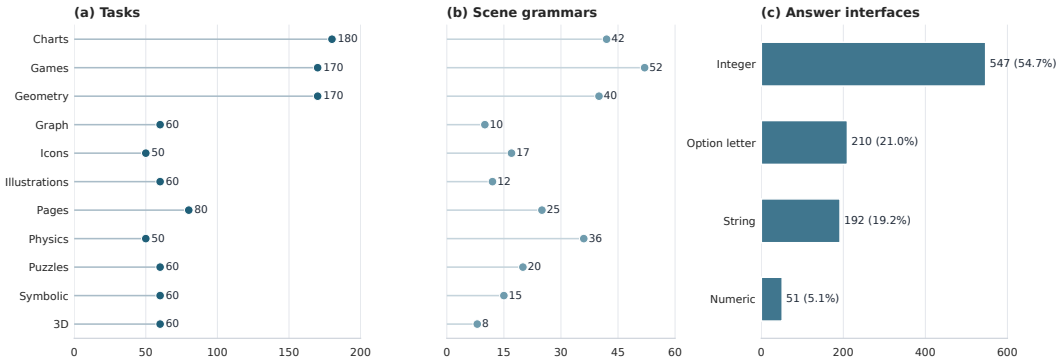


Figure 4: Composition of TRACE across 11 visual domains. The panels report the number of task programs and scene grammars in each domain, together with the overall distribution of answer interfaces.

4.2 Environment composition

TRACE comprises 1,000 tasks defined over 277 scene grammars spanning 11 visual domains. The domains organize related visual structures, rendering conventions, and object vocabularies for reporting and balancing; they are not themselves reasoning categories or sampling units. Figure 4 summarizes the distribution of tasks, scene grammars, and answer interfaces across domains. Figure 1 provides representative renderings, and Appendix C presents a domain-stratified task atlas.

The 277 scene grammars support a median of three tasks each (mean 3.61, range 1–26), indicating that visual construction is reused across multiple reasoning objectives rather than duplicated for each task. The task inventory includes integer, canonical numeric, string, and option-letter outputs; 790 tasks use open-form integer, numeric, or string answers.

Of the 1,000 tasks, 317 expose more than one internal query, producing 1,475 query variants in total. Because task selection precedes query selection, adding bounded query variants does not increase the sampling probability of the underlying task. The sampling distribution therefore reflects the authored task programs rather than the number of prompt-level variants available to each one.

Visual diversity alone does not establish diversity of computation. We therefore project the task inventory onto 13 compositional operation families, whose distribution across domains is shown in Figure 5. These families are an analytical view of the task programs rather than an additional taxonomy level. Assignments are multi-label because a composed program may require several operations; every task belongs to at least one family, and *direct retrieval* is assigned only when no other operation materially determines the answer. Appendix A.2 defines the complete set of families.

4.3 Controlled visual realization

Instance variation occurs at two distinct levels. Semantic controls alter the state on which the task program operates, including dimensions, cardinalities, values, thresholds, sequence lengths, and candidate sets. Prompt templates draw their dynamic labels and values from that same state, ensuring that the question and answer remain synchronized with task execution. All sampled choices are retained in the instance trace.

Once semantic execution fixes the state, query, and answer, scene-compatible realization controls vary how that state is presented. These controls include placement, panel arrangement, camera, palette, typography, materials, non-answer context, and bounded raster effects. A realization control is admitted only when it preserves the relations queried by the task program and therefore leaves the typed answer unchanged. Appendix A formalizes these variation layers and their required invariants.

For information-rich scenes, TRACE may additionally introduce curated context outside answer-bearing regions. Such context is constrained by scene-specific fit, collision, and contrast checks so that visual complexity can increase without modifying the underlying task state.

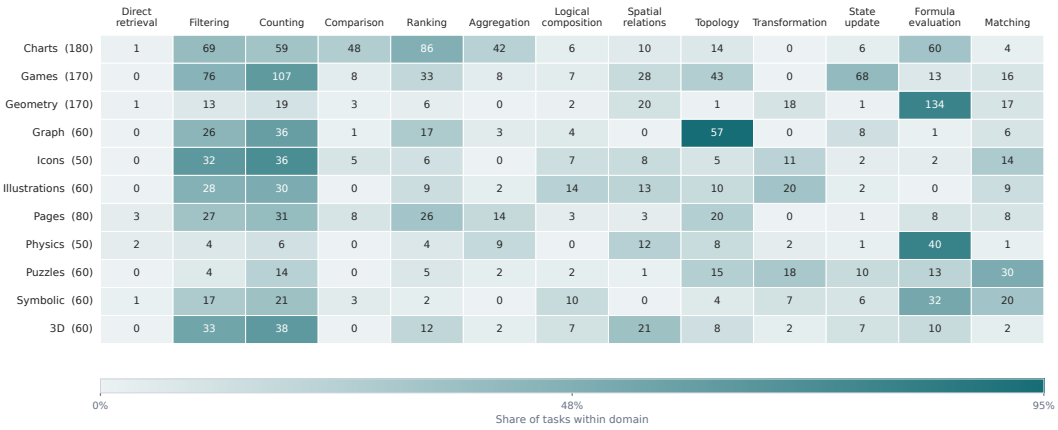


Figure 5: Distribution of task-program operations across visual domains. Cell labels report the number of tasks assigned to each operation family, while color indicates the corresponding within-domain share. Assignments are exhaustive and multi-label.

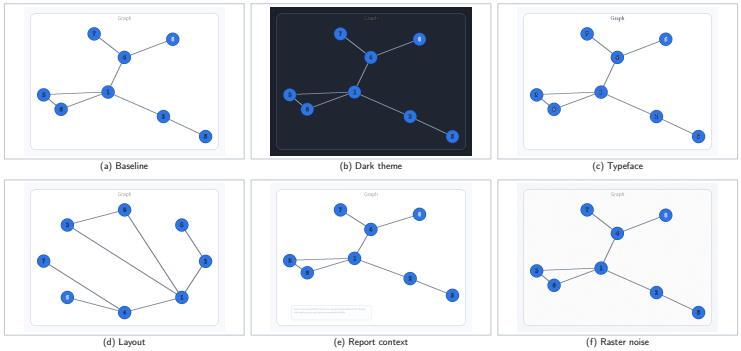


Figure 6: Answer-preserving realizations of a fixed semantic graph. Theme, typeface, layout, non-answer context, and raster treatment vary while the prompt, topology, query, and typed answer remain unchanged.

4.4 Instance validation

An instance is retained only if task execution produces a unique typed answer, the prompt and rendered image remain consistent with the semantic state, and deterministic replay reconstructs the same instance. Scene-specific rendering checks additionally test visibility, contrast, fit, and collisions.

These structural checks cannot guarantee that every generated question is semantically clear or that every valid rendering is equally legible. We therefore also inspect samples from every task for semantic clarity, prompt–image consistency, and visual legibility, complementing automatic validation with direct review of the rendered instances.

5 Experimental Setup

Our experiments test whether RLVR on procedurally generated TRACE instances improves visual reasoning beyond the distributions represented by the training tasks. We initialize from Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct [2] and apply the same training procedure at both model scales. Each resulting checkpoint is compared with its corresponding base checkpoint under a common external-evaluation protocol.

5.1 Training data and reward

Both model scales are trained on the same fixed set of 64,000 instances, comprising 64 independently generated examples from each of the 1,000 TRACE tasks. A separate held-out TRACE validation set contains two previously unseen instances per task, for 2,000 examples in total. All reported results use the final checkpoint after 500 updates.

The model may produce an unrestricted reasoning trace, but its response must terminate with a JSON object of the form `{"answer": ...}`. Let $R_a \in \{0, 1\}$ indicate whether the submitted answer exactly matches the target after type-aware canonicalization, and let $R_f \in \{0, 1\}$ indicate whether the response ends with a valid JSON object containing exactly the requested key. The reward is

$$R = 0.95R_a + 0.05R_f. \tag{8}$$

Answer correctness therefore determines the reward, while the smaller format term encourages a consistently parseable output interface.

5.2 Optimization

We optimize both models with group relative policy optimization (GRPO). Each update samples 128 prompts and generates eight responses per prompt. Rewards are normalized within each eight-response group and optimized using a clipped token-level policy objective. The 500 updates correspond to one shuffled pass over the 64,000 training prompts and produce 512,000 sampled responses in total.

We use full-parameter training with BF16 parameters, a learning rate of 10^{-6}, and one policy epoch per update. Rollout sampling uses temperature 1.0 and top-p 1.0. Training is performed on eight H100 80GB GPUs and takes 12.2 hours for the 3B model and 13.9 hours for the 7B model. The two scales share the same data, reward, sampling schedule, and optimization hyperparameters; Appendix B provides the complete configuration.

5.3 External evaluation

To measure transfer beyond the TRACE task distributions, we evaluate each base and trained checkpoint on 24 external benchmarks spanning six analysis groups: charts and tables, visual mathematics, science and general reasoning, spatial reasoning, perception and counting, and puzzles and logic. Each group contains four benchmarks. Table 2 reports the complete per-benchmark results, while Appendix Table 7 lists the evaluated splits, sample counts, and benchmark-specific references.

We evaluate each designated split in full, yielding 32,805 examples per model and decoding seed. Each benchmark retains its native evaluation metric. Category scores are computed as unweighted means of the four benchmarks in that group, and the overall score is the unweighted mean across all 24 benchmarks. This macro-averaging prevents benchmarks with larger evaluation sets from dominating the aggregate result.

For each fixed checkpoint, we repeat decoding with three seeds and report the mean and sample standard deviation. The seeds affect only response generation; the model weights remain unchanged. All checkpoints are evaluated with the same benchmark prompts, image-preprocessing bounds, parsers, scorers, and VLMEvalKit version [7]. Exact inference settings are provided in Appendix B.

At the 7B scale, we additionally evaluate publicly available synthetic-data RLVR checkpoints from Game-RL, Sphinx, and PC-GRPO [1, 17, 38]. We report Vero [29] separately as a real-image reference because its 600K-example training mixture draws from 59 datasets and includes real-image data, whereas our runs use 64K procedurally generated instances. All checkpoints are evaluated under the same external protocol. Because their training data, optimization, and compute are not matched, these comparisons characterize checkpoint outcomes rather than isolate the effect of a specific training method or data source.

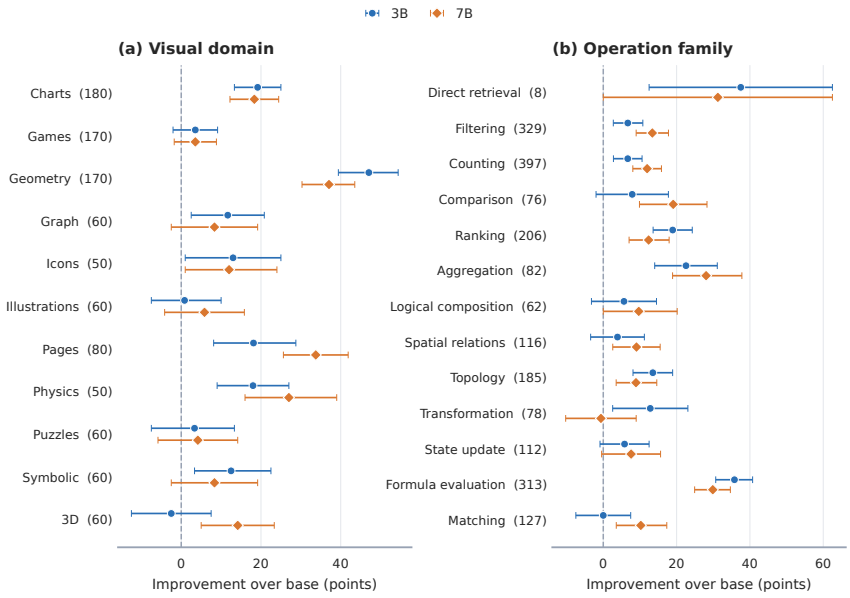


Figure 7: Held-out TRACE accuracy gains by visual domain (left) and operation family (right). Whiskers show 95% task-cluster bootstrap intervals; parentheses give task counts, and operation-family assignments overlap.

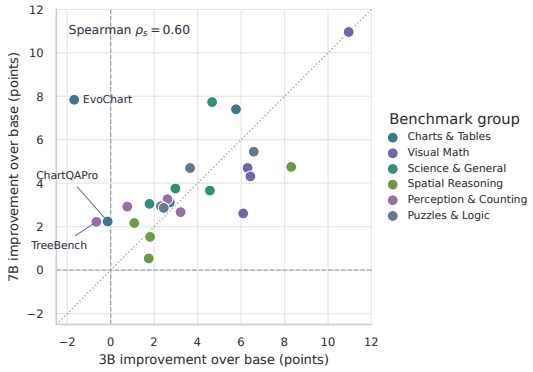


Figure 8: Benchmark gains at 3B versus 7B. Colors denote benchmark groups; dashed axes mark zero, the dotted diagonal marks equal gains, and $\rho_s = 0.60$.

6 Results

6.1 TRACE validation

RLVR substantially improves performance on unseen instances generated from the same task programs. Accuracy increases from 24.45 to 41.05 at 3B and from 34.25 to 51.55 at 7B, corresponding to gains of 16.60 and 17.30 percentage points, respectively. These results measure generalization to new semantic and visual realizations within the TRACE task distributions; the external benchmarks in Section 6.2 test whether the gains extend beyond those distributions.

The improvements are distributed across domains and computational structures rather than being driven by a single task family. Formula evaluation and aggregation show the largest operation-family gains at both model scales, whereas the 3D domain, transformation, and matching exhibit stronger scale-dependent behavior. Performance also improves across all four answer interfaces and for both single-query and multi-query tasks. Numeric-answer tasks show the largest increase, although this interface contains only 51 tasks, while option-letter tasks show the smallest. Appendix Table 6 reports these overlapping slices using two instances per task and one decoding seed.

Table 2: Per-benchmark transfer on the 24-benchmark external evaluation suite. Each score reports the mean and sample standard deviation over three decoding seeds, in percent. Smaller green/red subscripts in the TRACE columns give the change from the corresponding base model. Bold marks the best score among the base model and synthetic-data RLVR checkpoints at each model scale. Category and overall averages are unweighted across benchmarks.

	7B base + synthetic RLVR					Real-image	3B base + synthetic RLVR	
Benchmark	Base	TRACE	Game-RL	Sphinx	PC-GRPO	Vero	Base	TRACE
Charts & Tables								
ChartQAPro	45.8±0.2	**48.1±0.4**$_{+2.2}$	45.4±0.4	43.9±2.3	39.3±1.2	40.9±0.4	**31.6±0.7**	31.4±1.3$_{-0.1}$
CharXivReason	39.7±1.2	**47.1±0.4**$_{+7.4}$	39.7±1.0	40.7±0.4	41.2±0.7	45.8±0.8	28.9±1.1	**34.7±1.5**$_{+5.8}$
TableVQABench	75.2±0.9	**78.3±0.2**$_{+3.1}$	75.6±0.8	74.9±1.7	72.3±2.5	78.5±0.5	69.3±0.9	**72.0±0.3**$_{+2.7}$
EvoChart	57.1±0.7	**64.9±0.0**$_{+7.8}$	58.6±0.8	58.4±0.2	58.8±0.6	63.5±0.2	**48.5±0.5**	46.8±1.4$_{-1.7}$
Category average	54.5±0.3	**59.6±0.2**$_{+5.1}$	54.8±0.4	54.5±0.9	52.9±1.0	57.2±0.3	44.6±0.1	**46.2±0.9**$_{+1.7}$
Visual Math								
MathVision	24.8±0.7	**27.4±0.6**$_{+2.6}$	25.3±0.3	26.2±0.5	25.3±1.0	28.1±0.1	19.3±0.7	**25.4±0.9**$_{+6.1}$
MathVista	68.7±0.4	**73.4±0.5**$_{+4.7}$	69.9±1.0	71.0±0.7	70.8±0.6	76.8±0.4	58.1±3.7	**64.4±2.1**$_{+6.3}$
MathVerse	43.4±0.9	**47.8±0.5**$_{+4.3}$	44.1±1.5	45.3±2.4	44.9±1.0	50.9±1.1	33.6±2.0	**40.0±1.5**$_{+6.4}$
WeMath	35.2±2.9	**46.2±1.3**$_{+11.0}$	38.2±1.0	36.3±0.5	39.0±1.4	46.9±0.4	17.9±1.2	**28.8±0.4**$_{+10.9}$
Category average	43.0±0.8	**48.7±0.1**$_{+5.6}$	44.4±0.3	44.7±0.3	45.0±0.4	50.7±0.2	32.2±1.7	**39.7±0.8**$_{+7.4}$
Science & General								
PhyX mini MC	41.0±3.6	**48.7±0.8**$_{+7.7}$	41.4±4.5	42.4±3.3	42.0±2.9	46.6±0.2	32.8±10.0	**37.5±5.6**$_{+4.7}$
MMMU-ProVis	35.7±0.4	**39.4±0.7**$_{+3.7}$	36.1±0.6	37.6±0.2	36.9±1.2	39.7±0.5	26.6±0.3	**31.2±1.2**$_{+4.6}$
RealWorldQA	65.4±0.9	**68.5±0.7**$_{+3.1}$	66.0±0.3	67.3±0.2	67.8±1.0	69.8±0.7	60.3±0.4	**62.1±1.2**$_{+1.8}$
MMStar	61.9±0.9	**65.6±0.3**$_{+3.8}$	62.6±1.5	61.8±0.5	62.5±0.8	65.8±0.3	52.3±0.5	**55.2±1.0**$_{+3.0}$
Category average	51.0±0.8	**55.6±0.4**$_{+4.5}$	51.5±1.5	52.2±0.9	52.3±0.9	55.5±0.3	43.0±2.5	**46.5±1.8**$_{+3.5}$
Spatial Reasoning								
EmbSpatial	69.4±0.9	71.0±0.7$_{+1.5}$	69.5±0.9	70.9±0.7	**72.7±0.2**	70.6±0.1	59.1±1.0	**60.9±1.1**$_{+1.8}$
SpatialVizBench	35.1±0.1	**35.7±0.3**$_{+0.5}$	32.7±2.3	33.8±1.5	30.8±1.8	32.1±0.5	30.1±1.2	**31.8±1.5**$_{+1.8}$
CV-Bench 3D	76.2±2.0	**81.0±0.4**$_{+4.8}$	76.2±0.8	79.8±1.0	80.2±0.5	83.1±0.6	58.7±8.3	**67.0±3.7**$_{+8.3}$
ERQA	39.0±1.6	**41.2±1.3**$_{+2.2}$	38.8±1.7	40.3±2.0	40.4±0.8	44.6±0.8	35.3±0.8	**36.4±0.5**$_{+1.1}$
Category average	55.0±0.8	**57.2±0.3**$_{+2.2}$	54.3±1.0	56.2±1.0	56.0±0.5	57.6±0.3	45.8±2.4	**49.0±0.9**$_{+3.2}$
Perception & Counting								
BLINK	53.3±1.2	56.6±0.7$_{+3.3}$	53.8±1.2	55.7±1.1	**56.6±0.5**	57.5±0.7	44.5±2.1	**47.1±0.7**$_{+2.6}$
CountBenchQA	82.1±1.5	**84.8±1.4**$_{+2.7}$	82.6±1.4	83.3±1.3	80.6±1.2	83.0±0.2	65.4±1.6	**68.7±0.8**$_{+3.2}$
CountQA	19.6±1.2	**22.5±0.9**$_{+2.9}$	19.9±1.0	21.4±0.2	17.2±0.4	23.2±0.3	14.9±1.8	**15.6±0.6**$_{+0.8}$
TreeBench	38.0±1.4	40.2±2.1$_{+2.2}$	37.9±1.0	39.8±1.8	**40.4±2.0**	41.5±0.9	**39.3±2.0**	38.6±1.2$_{-0.7}$
Category average	48.3±0.3	**51.0±0.4**$_{+2.8}$	48.5±0.2	50.0±0.2	48.7±0.2	51.3±0.0	41.0±0.5	**42.5±0.3**$_{+1.5}$
Puzzles & Logic								
PuzzleVQA	44.6±0.6	**50.0±0.5**$_{+5.5}$	40.1±1.5	48.6±1.1	47.9±1.1	49.2±0.2	32.6±1.2	**39.1±1.1**$_{+6.6}$
VisualPuzzles	31.1±0.1	**34.0±0.8**$_{+2.9}$	29.4±0.7	33.2±0.9	33.1±1.7	37.0±0.4	26.2±2.7	**28.5±0.7**$_{+2.3}$
LogicVista	42.7±2.2	**47.4±0.7**$_{+4.7}$	43.3±0.9	44.7±1.2	43.0±0.9	47.7±0.7	36.5±1.2	**40.1±1.7**$_{+3.7}$
MME-Reasoning	25.2±1.2	**28.0±0.9**$_{+2.9}$	26.1±1.1	27.1±0.3	27.5±1.0	30.6±0.7	22.6±1.8	**25.1±1.5**$_{+2.4}$
Category average	35.9±0.7	**39.9±0.5**$_{+4.0}$	34.7±0.7	38.4±0.1	37.9±0.8	41.1±0.1	29.5±0.6	**33.2±0.9**$_{+3.7}$
Overall average	47.9±0.3	**52.0±0.2**$_{+4.1}$	48.0±0.4	49.3±0.2	48.8±0.3	52.2±0.1	39.3±0.6	**42.9±0.4**$_{+3.5}$

6.2 External benchmark transfer

Table 2 reports transfer across the complete 24-benchmark evaluation suite.

Training on TRACE improves the benchmark-macro average at both model scales. At 3B, the score increases from 39.34 ± 0.63 to 42.85 ± 0.39, a gain of 3.51 percentage points. At 7B, it increases from 47.93 ± 0.30 to 51.99 ± 0.17, a gain of 4.06 points. The improvements are broad: all six category averages increase at both scales, mean performance improves on 21 of 24 benchmarks at 3B and on all 24 benchmarks at 7B, and paired item-bootstrap intervals exclude zero on 18 and 21 benchmarks, respectively. EvoChart at 3B is the only benchmark whose interval lies entirely below zero; the remaining intervals either exclude zero in the positive direction or include it. Appendix Table 8 reports the complete benchmark-level intervals.

Visual mathematics shows the largest category-level improvement at both scales, with gains of 7.44 points at 3B and 5.65 points at 7B. WeMath shows the largest individual improvement, increasing by 10.95 points at 3B and 10.96 points at 7B. The gains therefore do not arise solely from improved performance on held-out procedural images, but extend across external benchmarks with different visual formats, task structures, and evaluation metrics.

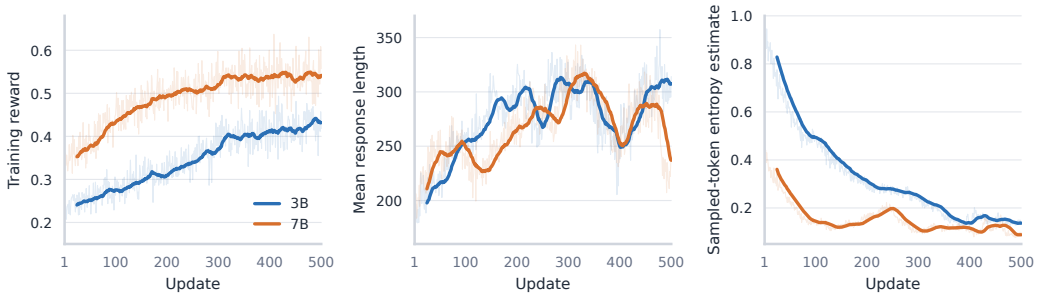


Figure 9: Optimization trajectories over 500 RLVR updates. Faint lines show per-update values, and solid lines show 25-update moving averages. Training reward is the bounded answer-and-format reward defined in Section 5. Response length is measured over sampled completions, and entropy is the sampled-token estimate recorded by the policy optimizer.

Under the common evaluation protocol, the TRACE-trained 7B checkpoint obtains a macro-average 3.94 points higher than Game-RL-7B, 2.64 points higher than Sphinx-7B, and 3.18 points higher than PC-GRPO-7B. Among these synthetic-data RLVR checkpoints, it also has the highest reported mean on 21 of 24 benchmarks and the highest category average in all six groups. These comparisons describe checkpoint outcomes rather than isolate the effect of a particular method or training distribution, because the systems are not matched in data, optimization, or compute. Table 2 reports Vero separately because its substantially larger training mixture includes real images.

Figure 8 shows moderate consistency in the relative transfer patterns across model scales. The negative mean changes at 3B on ChartQAPro (-0.14), TreeBench (-0.66), and EvoChart (-1.68) become gains of 2.24, 2.22, and 7.84 points at 7B, respectively. EvoChart is the largest departure from the equal-gain diagonal. Thus, although the overall transfer pattern is positively associated across scales, the benefit of a fixed procedural training mixture remains dependent on model capacity and benchmark demands.

6.3 Training dynamics

Training reward increases at both model scales despite substantial batch-to-batch variation. Averaged over the first and final 25 updates, mean reward rises from 0.240 to 0.432 at 3B and from 0.353 to 0.542 at 7B. The 7B model therefore begins and ends with higher reward, while both models exhibit continued improvement throughout the single training pass.

The scales differ more clearly in their response-length trajectories. Mean completion length increases from 198 to 308 tokens at 3B, compared with a smaller increase from 211 to 237 tokens at 7B. Over the same intervals, the sampled-token entropy estimate decreases from 0.829 to 0.137 at 3B and from 0.361 to 0.088 at 7B. Optimization therefore raises reward while making the sampled token distributions more concentrated, without reducing the models to shorter responses; both retain multi-hundred-token completions at the end of training.

7 Limitations and Future Work

The TRACE taxonomy is explicit but human-designed: task boundaries can still involve judgment, and the 1,000 authored tasks do not exhaust visual reasoning. The current operation families characterize computation types but do not define a common notion of difficulty across domains. Future work could model complexity through program structure, semantic-state size, visual density, and empirical solve rates.

Our experiments show transfer from the complete TRACE mixture but do not attribute gains to particular domains, operation families, sampling choices, or rendering controls. Such attribution would require mixture ablations. The current study also uses uniform task sampling over one fixed training pass; the explicit task identities and generation parameters could instead support difficulty-aware curricula, adaptive sampling, and mixture optimization based on model competence or learning progress.

Results are based on one training run per model scale, so variation across decoding seeds reflects inference variability rather than independent training variation. Comparisons with released RLVR checkpoints are descriptive because their data, optimization, and compute are unmatched. The external suite is broad but cannot establish transfer to every natural-image distribution, and some benchmarks use model-based graders. Finally, procedural consistency does not guarantee equal perceptual difficulty: synthetic renderers may introduce regularities or occasional ambiguities not found in natural images. Richer rendering sources and replay-based failure analysis are natural directions for future work.

8 Conclusion

TRACE formulates procedural visual data as an executable environment in which scene grammars, task programs, bounded queries, and reward contracts are explicit and separable. This structure provides stable units for generation, task-balanced sampling, verification, and analysis while keeping every instance exactly verifiable and deterministically replayable across semantic and visual variation. RLVR on 64,000 TRACE instances improves the macro-average across 24 external benchmarks by 3.51 points at 3B and 4.06 points at 7B, with positive mean changes on 21 of 24 and all 24 benchmarks, respectively. These results provide evidence that broad, structured procedural experience can support transferable visual reasoning rather than only improving performance on regenerated instances from the same task distributions.

Broader Impact

TRACE may support more reproducible multimodal-reasoning research by making task definitions, supervision, and reward computation inspectable, while reducing reliance on sensitive or copyrighted source imagery in settings where procedural data are suitable. At the same time, stronger visual-reasoning systems may be used in surveillance, automated assessment, or other consequential decision pipelines, and large-scale RLVR training can require substantial computational resources. Releases built on TRACE should therefore document intended uses, licensing, resource requirements, and downstream-use constraints.

References

[1] Md Tanvirul Alam, Saksham Aggarwal, Justin Yang Chae, and Nidhi Rastogi. SPHINX: A synthetic environment for visual perception and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9489–9499, 2026.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Jiangjie Chen, Qianyu He, Siyu Yuan, Aili Chen, Zhicheng Cai, Weinan Dai, et al. Enigmata: Scaling logical reasoning in large language models with synthetic verifiable puzzles. *arXiv preprint arXiv:2505.19914*, 2025.

[4] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

[5] Yew Ken Chia, Vernon Toh Yan Han, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. PuzzleVQA: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. *arXiv preprint arXiv:2403.13315*, 2024.

[6] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

[7] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024.

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

[9] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024.

[10] Google DeepMind. Embodied reasoning question answer (ERQA) benchmark. GitHub repository, 2025. URL `https://github.com/embodiedreasoning/ERQA`.

[11] Xinyi Gu, Jiayuan Mao, Zhang-Wei Hong, Zhuoran Yu, Pengyuan Li, Dhiraj Joshi, Rogerio Feris, and Zexue He. Composition-grounded data synthesis for visual reasoning. *arXiv preprint arXiv:2510.15040*, 2025.

[12] Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. OpenThoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

[13] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[14] Muye Huang, Han Lai, Xinyu Zhang, Wenjun Wu, Jie Ma, Lingling Zhang, and Jun Liu. EvoChart: A benchmark and a self-training approach towards real-world chart understanding. *arXiv preprint arXiv:2409.01577*, 2025.

[15] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, et al. Vision-R1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

[16] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[17] Ahmadreza Jeddi, Hakki Can Karaimer, Hue Nguyen, Zhongling Wang, Ke Zhao, Javad Rajabi, Ran Zhang, Raghav Goyal, Konstantinos G. Derpanis, Babak Taati, and Radek Grzeszczuk. PuzzleCraft: Exploration-aware curriculum learning for puzzle-based RLVR in VLMs. *arXiv preprint arXiv:2512.14944*, 2025. URL `https://arxiv.org/abs/2512.14944`.

[18] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[19] Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. TableVQA-Bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*, 2024.

[20] Shenshen Li, Xing Xu, Kaiyuan Deng, Lei Wang, Heng Tao Shen, and Fumin Shen. Truth in the few: High-value data selection for efficient multi-modal reasoning. *arXiv preprint arXiv:2506.04755*, 2025.

[21] Yiqing Liang, Jielin Qiu, Wenhao Ding, Zuxin Liu, James Tompkin, Mengdi Xu, Mengzhou Xia, Zhengzhong Tu, Laixi Shi, and Jiacheng Zhu. MoDoMoDo: Multi-domain data mixtures for multimodal LLM reinforcement learning. *arXiv preprint arXiv:2505.24871*, 2025.

[22] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

[23] 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. *arXiv preprint arXiv:2310.02255*, 2024.

[24] Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. ChartQAPro: A more diverse and challenging benchmark for chart question answering. *arXiv preprint arXiv:2504.05506*, 2025.

[25] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, et al. MM-Eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[26] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching CLIP to count to ten. *arXiv preprint arXiv:2302.12066*, 2023.

[27] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, et al. LMM-R1: Empowering 3b LMMs with strong reasoning abilities through two-stage rule-based RL. *arXiv preprint arXiv:2503.07536*, 2025.

[28] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gong Que, Shanglin Lei, Yifan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

[29] Gabriel Sarch, Linrong Cai, Qunzhong Wang, Haoyang Wu, Danqi Chen, and Zhuang Liu. Vero: An open RL recipe for general visual reasoning. *arXiv preprint arXiv:2604.04917*, 2026.

[30] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[31] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, et al. VLM-R1: A stable and generalizable R1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

[32] Hui Shen, Taiqiang Wu, Qi Han, Yunta Hsieh, Jizhou Wang, Yuyue Zhang, Yuxin Cheng, Zijian Hao, Yuansheng Ni, Xin Wang, Zhongwei Wan, Kai Zhang, Wendong Xu, Jing Xiong, Ping Luo, Wenhu Chen, Chaofan Tao, Zhuoqing Mao, and Ngai Wong. PhyX: Does your model have the wits for physical reasoning? *arXiv preprint arXiv:2505.15929*, 2025.

[33] Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. VisualPuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025.

[34] Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. Reasoning Gym: Reasoning environments for reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2505.24760*, 2025.

[35] Haoxiang Sun, Lizhen Xu, Bing Zhao, Wotao Yin, Wei Wang, Boyu Yang, Rui Wang, and Hu Wei. DeepVision-103K: A visually diverse, broad-coverage, and verifiable mathematical dataset for multimodal reasoning. *arXiv preprint arXiv:2602.16742*, 2026.

[36] Jayant Sravan Tamarapalli, Rynaa Grover, Nilay Pande, and Sahiti Yerramilli. CountQA: How well do MLLMs count in the wild? *arXiv preprint arXiv:2508.06585*, 2025.

[37] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-RFT: Reinforcement fine-tuning for visual reasoning of vision language models. *arXiv preprint arXiv:2503.20752*, 2025.

[38] Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, Jun Zhao, Yanbo Wen, Fan Song, Jiahao Zhan, Yuyang Lu, et al. Game-RL: Synthesizing multimodal verifiable game data to boost VLMs' general reasoning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=e4FqU4SyHL.

[39] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. *arXiv preprint arXiv:2406.16860*, 2024.

[40] Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, and Zhaoxiang Zhang. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. *arXiv preprint arXiv:2507.07999*, 2026.

[41] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. *arXiv preprint arXiv:2402.14804*, 2024.

[42] Siting Wang, Minnan Pei, Luoyang Sun, Cheng Deng, Yuchen Li, Kun Shao, Zheng Tian, Haifeng Zhang, and Jun Wang. SpatialViz-Bench: A cognitively-grounded benchmark for diagnosing spatial visualization in MLLMs. *arXiv preprint arXiv:2507.07610*, 2026.

[43] Xiyao Wang, Zhengyuan Yang, Chao Feng, Yongyuan Liang, Yuhang Zhou, Xiaoyu Liu, et al. ViCrit: A verifiable reinforcement learning proxy task for visual perception in VLMs. *arXiv preprint arXiv:2506.10128*, 2025.

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

[45] Zifu Wang, Junyi Zhu, Bo Tang, Zhiyu Li, Feiyu Xiong, Jiaqian Yu, and Matthew B. Blaschko. Jigsaw-R1: A study of rule-based visual reinforcement learning with jigsaw puzzles. *arXiv preprint arXiv:2505.23590*, 2025.

[46] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. *arXiv preprint arXiv:2406.18521*, 2024.

[47] Zijian Wu, Jinjie Ni, Xiangyan Liu, Zichen Liu, Hang Yan, and Michael Qizhe Shieh. SynthRL: Scaling visual reasoning with verifiable data synthesis. *arXiv preprint arXiv:2506.02096*, 2025.

[48] xAI. RealWorldQA. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/xai-org/RealworldQA.

[49] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. LogicVista: Multimodal LLM logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

[50] Jie Yang, Feipeng Ma, Zitian Wang, Dacheng Yin, Kang Rong, Fengyun Rao, and Ruimao Zhang. WeThink: Toward general-purpose vision-language reasoning via reinforcement learning. *arXiv preprint arXiv:2506.07905*, 2025.

[51] Tianze Yang, Yucheng Shi, Ruitong Sun, Jingyuan Huang, Ninghao Liu, and Jin Sun. TRON: Targeted rule-verifiable online environments for visual reasoning RL. *arXiv preprint arXiv:2606.01599*, 2026. URL https://arxiv.org/abs/2606.01599.

[52] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, et al. R1-Onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

[53] Yida Yin, Harish Krishnakumar, Chung Peng Lee, Boya Zeng, Wenhao Chai, Shengbang Tong, Wenhu Chen, Hu Xu, Xingyu Fu, Gabriel Sarch, Aleksandra Korolova, and Zhuang Liu. Worldbench: A challenging and visually diverse multimodal reasoning benchmark. *arXiv preprint arXiv:2606.06538*, 2026.

[54] Jiakang Yuan, Tianshuo Peng, Yilei Jiang, Yiting Lu, Renrui Zhang, Kaituo Feng, Chaoyou Fu, Tao Chen, Lei Bai, Bo Zhang, and Xiangyu Yue. MME-Reasoning: A comprehensive benchmark for logical reasoning in MLLMs. *arXiv preprint arXiv:2505.21327*, 2025.

[55] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2025.

[56] Yuheng Zha, Kun Zhou, Yujia Wu, Yushu Wang, Jie Feng, Zhi Xu, Shibo Hao, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. Vision-G1: Towards general reasoning vision-language models via reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (33):28131–28139, 2026. doi: 10.1609/aaai.v40i33.40039.

[57] Jieyu Zhang, Weikai Huang, Zixian Ma, Oscar Michel, Dong He, Tanmay Gupta, Wei-Chiu Ma, Ali Farhadi, Aniruddha Kembhavi, and Ranjay Krishna. Task me anything. *arXiv preprint arXiv:2406.11775*, 2024.

[58] 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? *arXiv preprint arXiv:2403.14624*, 2024.

A Environment Construction and Validation

A.1 Domain construction profiles

Prompts are assembled from scene-, task-, and query-level templates. Dynamic labels, values, and answers are derived from the same semantic state and task execution recorded in the instance trace. The profiles below therefore summarize the domain-specific state representations, rendering conventions, and visible vocabularies; the common generation and reward pipeline is described in Section 4.

Charts. Chart scenes are constructed from procedural marks, axes, scales, legends, and panels spanning Cartesian, radial, geographic, flow, tabular, and compositional layouts. Rendering controls vary chart treatment, palette, typography, tick density, mark style, and compatible report context without modifying the underlying data. Curated entity, temporal, scientific, and compact label pools provide distinct series, category, panel, and region names within each figure.

Games. Game scenes combine procedural boards, tracks, hands, scores, pieces, enemies, and action states with curated cards, dice, tokens, and sprite-like assets. Each renderer follows the layout and interaction rules of its scene grammar rather than a shared generic board template. Coordinates, score displays, status labels, and action cues are derived from the simulated state, while optional interface overlays and surface treatments provide visual variation.

Geometry. Geometry scenes are generated from analytic constructions of points, lines, angles, polygons, circles, solids, coordinate systems, and measurement instruments under solved numeric constraints. Textbook, whiteboard, and technical-diagram profiles vary stroke, shading, notation placement, and typography. Point labels, measurements, equations, and answer options are bound after construction so that the visible diagram and symbolic problem statement remain consistent with the same geometric state.

Graphs. Graph scenes begin with explicit nodes, edges, weights, paths, and state transitions, which are then realized through circular, layered, tree, route, or force-directed layouts. Node shape, edge style, palette, and framing vary independently of topology. Numeric, alphabetic, and named node vocabularies are sampled without collisions, and labels and edge values remain attached to their semantic elements across layout changes.

Icons. Icon scenes populate fields, grids, paths, rings, paired canvases, and rendered option sets with curated icon assets. Procedural controls vary icon identity, color, scale, rotation, spacing, ordering, and transformation while preserving legibility and answer uniqueness. Marker and option labels are placed as part of the scene geometry, allowing counting, adjacency, transformation, and matching tasks to share a consistent visual vocabulary.

Illustrations. Illustration scenes compose semantic object, part, and tile catalogs into rooms, construction sites, libraries, tactical maps, and isometric environments. Scene-specific layout rules determine occupancy, overlap, depth, and permitted object relations, after which palette, asset, framing, and surface treatments vary the appearance. Explicit tile, object, and option labels are introduced only when required by the task program.

Pages. Page scenes represent structured records through forms, tables, calendars, schedules, application controls, documents, and infographic components. Templates govern visual hierarchy, alignment, interface framing, spacing, and typography, while information-density and context profiles vary the presentation. Field names, section headings, dates, values, and control labels remain linked to the underlying document state.

Physics. Physics scenes assemble apparatus, forces, wires, rays, fields, graphs, and measurement instruments from explicit physical quantities and component states. Technical-diagram profiles vary line work, component treatment, background, and label placement while preserving the solved configuration. Quantities, units, directions, and component names are rendered from the same state used by formula-evaluation, comparison, and state-update programs.

Table 3: Variation layers and the invariants they preserve.

Layer	Representative controls	Required invariant
Semantic instance	Dimensions, cardinalities, values, operation length, candidate set	Scene grammar and task program
Structural realization	Placement, spacing, panel arrangement, framing, answer-preserving camera	Semantic relations and answer
Surface and context	Palette, typography, skins, materials, labels, non-answer text	Semantic state and answer
Post-render	Blur, resampling, compression, tone, noise, texture	Canvas geometry and answer

Puzzles. Puzzle scenes include rule-governed grids, mazes, pipes, matrices, nets, voxel boards, and candidate sets sampled under validity and uniqueness constraints. Symbols, board skins, spacing, missing-cell treatments, and option scaffolds vary the presentation without revealing the underlying rule. Coordinates, candidate labels, and compact clues are added only when required by the puzzle definition or answer interface.

Symbolic. Symbolic scenes use notation-specific generators for clocks, abaci, dice, spinners, musical staves, logic circuits, formal tables, tapes, and encoded glyph systems. Symbols and transitions are represented as executable state rather than decorative text. Fonts, notation cards, palettes, spacing, and layout may vary, while tokens, values, note names, and state labels remain bound to the formal structure being queried.

Three-dimensional scenes. Three-dimensional scenes arrange procedural meshes and object assets under explicit camera, projection, depth, occlusion, and surface constraints. Viewpoint, material, lighting, ground or wall treatment, and framing vary without changing the queried spatial relation. Object, lane, wall, point, and candidate labels are placed after projection and checked against the final image geometry for visibility and separation.

A.2 Operation-family definitions

Figure 5 classifies tasks according to the operations that materially contribute to their outputs. Assignments are multi-label when a task program composes several operations.

Direct retrieval denotes a localized visible lookup for which no other substantive operation determines the answer. *Filtering* constructs a subset using a predicate, relation, or named scope, and *counting* returns the cardinality of such a set. *Comparison* evaluates equality, order, thresholds, or intervals, whereas *ranking* orders candidates or selects an extremum or ordinal item. *Aggregation* reduces a collection using a sum, mean, median, cumulative total, share, mass, or related statistic.

Logical composition combines predicates or sets through conjunction, disjunction, negation, union, intersection, or difference. *Spatial relations* covers positional, directional, metric, overlap, containment, and occlusion relations. *Topology* covers connectivity, paths, components, cycles, traversal, and reachability. *Matching* evaluates equivalence, correspondence, consistency, completion, or candidate-to-reference agreement.

Transformation applies or infers a geometric, symbolic, representational, folding, projection, overlay, or reconstruction operation. *State update* applies a move, action, counterfactual edit, or discrete transition. *Formula evaluation* uses arithmetic, algebra, a domain-specific formula, a derived metric, or a numeric difference, ratio, or rate.

A.3 Invariant enforcement

Semantic generation first enforces task-specific constraints and answer uniqueness. Structural realization is completed before geometry-dependent verifier state is projected into image coordinates, and post-render effects preserve the final canvas coordinate system. Render-only controls may not modify task-program operands, answer values, or the reward contract.

Independent seeded streams and the recorded instance trace allow the semantic state, prompt, rendering choices, and verifier state to be replayed together. The precise ordering of optional finishing layers remains renderer-specific. Table 3 summarizes the variation levels and their required invariants.

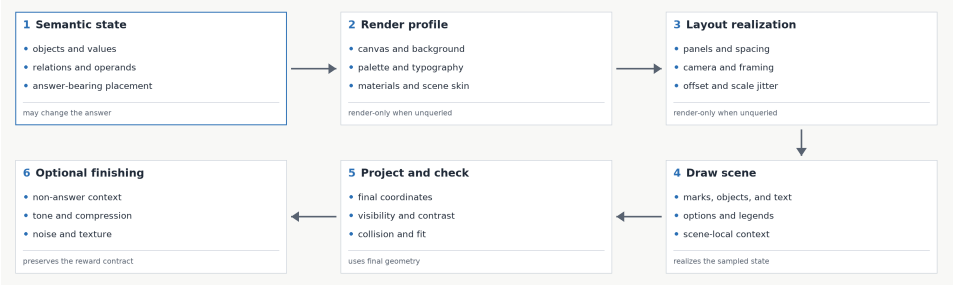


Figure 10: Control flow for visual realization. Structural and surface controls are treated as render-only when they preserve the queried semantic relations and typed answer. Answer-bearing geometry is projected before validation, while optional context and raster finishing remain scene-specific and are recorded in the instance trace.

Table 4: Controlled profiles in the fixed-instance rendering sweep.

Profile	Controlled axis	Realization
Baseline	Reference	Light neutral theme, sans-serif type, fixed layout, clean background
Dark theme	Theme	Dark publication treatment and analytics palette; topology and layout retained
Typeface	Typography	EB Garamond node-label typeface; baseline palette and layout retained
Layout	Structure	Circular node placement replaces the baseline spring layout
Report context	Non-answer context	Curated paragraph context in reserved non-answer panel space
Raster noise	Post-render	Gaussian noise with fixed sigma; coordinate geometry retained

Figure 10 illustrates how semantic execution, structural realization, projection, and surface treatment interact while preserving the task output.

Table 4 specifies the controlled realizations used in Figure 6.

B Experimental Details

Table 5 reports the configuration shared by the 3B and 7B training runs.

External inference uses temperature 0.6, top-p 1.0, a 4,096-token response limit, and decoding seeds 42, 43, and 44. The seeds affect response generation only; model weights remain fixed across decoding runs.

B.1 TRACE validation slices

Table 6 reports held-out TRACE accuracy by answer interface and query structure. Each task contributes two previously unseen instances, and all slice scores use one decoding seed. The rows are overlapping analytical views of the same task inventory rather than separate evaluation sets.

B.2 External evaluation suite

Each benchmark is evaluated with its standard scorer when available. A small number of open-answer benchmarks retain their fixed model-based grading procedures; the same grader, prompt, and parser are used for every checkpoint. Table 7 lists the complete benchmark suite, references, and evaluated sample counts.

B.3 External benchmark uncertainty

Table 8 reports paired confidence intervals for the improvement from each base checkpoint to its TRACE-trained counterpart. For each bootstrap replicate, evaluation items are resampled jointly with the predictions from all three decoding seeds. Benchmarks with grouped metrics are resampled at their official scoring unit: problem families for WeMath and items within official splits for TableVQABench.

Table 5: RLVR configuration shared across the 3B and 7B runs.

Setting	Value
Training data	64,000 prompts; 1,000 tasks
Image pixel cap	1,280,000 per image
Updates	500; one shuffled data pass
Validation	2,000 instances every 100 updates
Prompt batch / rollouts	128 / 8
Sampled responses	512,000
Prompt / response caps	2,048 / 2,048 tokens
Optimization scope	Full model; vision encoder trainable; BF16
Actor optimizer	AdamW; LR 10^{-6}; weight decay 0.01
Policy epochs	1
Policy clipping	0.2 lower; 0.3 upper; dual 3.0
Rollout sampling	temperature 1.0; top-p 1.0
Reward	0.95 exact answer + 0.05 JSON format
Hardware	8 $\times$ H100 80GB
Runtime	12.2 h (3B); 13.9 h (7B)

Table 6: Accuracy by answer interface and query structure on the 2,000-instance TRACE validation set. Each task contributes two unseen instances and scores use one decoding seed. Rows are analytical slices of the same task inventory.

		3B			7B		
Slice	Tasks	Base	After RLVR	Δ	Base	After RLVR	Δ
Answer interface							
Integer	547	22.39	37.84	+15.45	30.35	49.36	+19.01
Numeric	51	24.51	95.10	+70.59	49.02	94.12	+45.10
Option letter	210	26.43	30.95	+4.52	33.33	40.48	+7.14
String	192	28.12	46.88	+18.75	42.45	58.59	+16.15
Query structure							
Single-query tasks	683	24.52	39.82	+15.30	33.53	50.88	+17.35
Multi-query tasks	317	24.29	43.69	+19.40	35.80	53.00	+17.19

C Representative Task Atlas

Each page presents twelve seeded examples from one visual domain. Panels pair each rendered instance with its question and typed answer. Rule-heavy questions are condensed only when required for legibility.

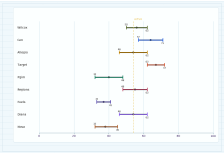
Table 7: External evaluation suite. Each model and decoding seed is scored on 32,805 examples.

Category	Benchmark	Rows
Charts & Tables	ChartQAPro [24]	1,948
	CharXivReason [46]	1,000
	TableVQABench [19]	1,500
	EvoChart [14]	1,250
Visual Math	MathVision [41]	3,040
	MathVista [23]	1,000
	MathVerse [58]	788
	WeMath [28]	1,740
Science & General	PhyX mini MC [32]	1,000
	MMMU-ProVis [55]	1,730
	RealWorldQA [48]	765
	MMStar [4]	1,500
Spatial Reasoning	EmbSpatial [6]	3,640
	SpatialVizBench [42]	1,180
	CV-Bench 3D [39]	1,200
	ERQA [10]	400
Perception & Counting	BLINK [9]	1,901
	CountBenchQA [26]	487
	CountQA [36]	1,528
	TreeBench [40]	405
Puzzles & Logic	PuzzleVQA [5]	2,000
	VisualPuzzles [33]	1,168
	LogicVista [49]	447
	MME-Reasoning [54]	1,188

Table 8: Paired Base-to-TRACE improvements on the external evaluation suite, in percentage points. Intervals are 95% bootstrap percentile intervals, computed from the 2.5th and 97.5th percentiles of 10,000 paired item-bootstrap replicates; each sampled item carries results from all three decoding seeds. Bold intervals exclude zero. These intervals quantify variation over evaluation items rather than independently trained checkpoints.

	3B		7B	
Benchmark	Δ	95% CI	Δ	95% CI
Charts & Tables				
ChartQAPro	-0.14	$[-1.30, +1.06]$	$+2.24$	$[+0.76, +3.66]$
CharXivReason	$+5.77$	$[+4.13, +7.40]$	$+7.40$	$[+5.33, +9.50]$
TableVQABench	$+2.72$	$[+1.78, +3.69]$	$+3.11$	$[+2.01, +4.21]$
EvoChart	-1.68	$[-3.20, -0.16]$	$+7.84$	$[+5.89, +9.79]$
Visual Math				
MathVision	$+6.10$	$[+5.02, +7.18]$	$+2.61$	$[+1.54, +3.66]$
MathVista	$+6.30$	$[+4.53, +8.07]$	$+4.70$	$[+2.80, +6.53]$
MathVerse	$+6.43$	$[+4.27, +8.54]$	$+4.31$	$[+2.12, +6.56]$
WeMath	$+10.95$	$[+8.73, +13.24]$	$+10.96$	$[+8.39, +13.56]$
Science & General				
PhyX mini MC	$+4.67$	$[+3.40, +5.90]$	$+7.73$	$[+5.10, +10.37]$
MMMU-ProVis	$+4.57$	$[+3.12, +6.05]$	$+3.66$	$[+2.22, +5.13]$
RealWorldQA	$+1.79$	$[+0.92, +2.66]$	$+3.05$	$[+1.44, +4.71]$
MMStar	$+2.98$	$[+1.93, +4.04]$	$+3.76$	$[+2.44, +5.16]$
Spatial Reasoning				
EmbSpatial	$+1.81$	$[+1.34, +2.28]$	$+1.53$	$[+0.94, +2.09]$
SpatialVizBench	$+1.75$	$[-0.28, +3.84]$	$+0.54$	$[-1.64, +2.85]$
CV-Bench 3D	$+8.31$	$[+6.50, +10.11]$	$+4.75$	$[+3.11, +6.39]$
ERQA	$+1.08$	$[-0.50, +2.75]$	$+2.17$	$[-0.17, +4.50]$
Perception & Counting				
BLINK	$+2.61$	$[+1.75, +3.49]$	$+3.26$	$[+2.03, +4.54]$
CountBenchQA	$+3.22$	$[+1.85, +4.65]$	$+2.67$	$[+1.03, +4.31]$
CountQA	$+0.76$	$[-0.07, +1.59]$	$+2.92$	$[+1.44, +4.43]$
TreeBench	-0.66	$[-2.47, +1.15]$	$+2.22$	$[+0.00, +4.44]$
Puzzles & Logic				
PuzzleVQA	$+6.58$	$[+4.90, +8.23]$	$+5.45$	$[+3.85, +7.08]$
VisualPuzzles	$+2.31$	$[+0.37, +4.22]$	$+2.94$	$[+0.88, +4.97]$
LogicVista	$+3.65$	$[+0.45, +6.79]$	$+4.70$	$[+1.42, +7.98]$
MME-Reasoning	$+2.44$	$[+0.76, +4.07]$	$+2.86$	$[+1.04, +4.69]$

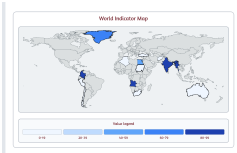
Error Interval: Reference Containment



The image shows a horizontal interval chart with labeled categories, point estimates, and lower-to-upper interval whiskers. How many intervals contain the reference value 54?

Answer: 4

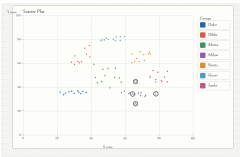
Region Map: Numeric Interval Region



This image shows a world map with selected countries colored by value and a color legend. How many countries fall in value bins between 20 and 99, inclusive?

Answer: 6

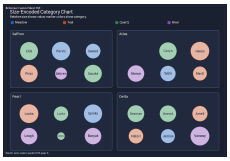
Scatter Cluster: Centroid Option Selection



The visual shows a scatter plot with several colored point clusters and a matching legend. Find the centroid of cluster "Milow" by eye. Which option marker is closest?

Answer: A

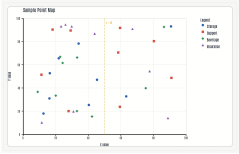
Size Encoding: Category Extremum Across Panels



The image shows multiple packed bubble-chart panels where bubble size indicates value and color indicates category. For category "River", which panel contains the item with the largest size?

Answer: Delta

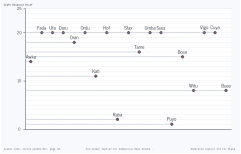
Scatter Points: Category Threshold Point



The chart shows a categorized scatter plot with individual data points, numeric x- and y-axes, and a legend. How many plotted points from category "Storage" satisfy x greater than 50?

Answer: 2

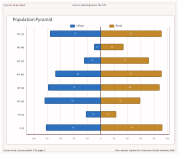
Single Series: Threshold Count



The image shows an ordered labeled dot plot where point labels identify y-values from left to right. Determine how many marks show values strictly greater than 19.

Answer: 10

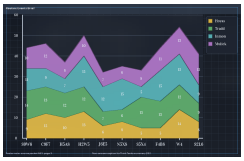
Population Pyramid: Age Group Threshold



The image shows a mirrored horizontal bar chart with one row per age group. The left and right bars show the two legend series on the same positive scale. In how many age groups is the sum of Urban and Rural at most 140?

Answer: 3

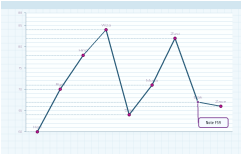
Area: Stacked Band Dominance



The visual shows one stacked area chart with colored category bands, numeric band values, and a legend. Sum each category's band values over "S9W6" through "H2W5" in displayed x-axis order; which category is largest?

Answer: Tradd

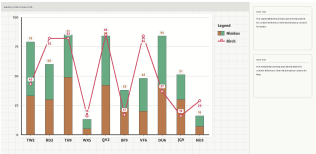
Annotated Series: Callout Endpoint Change



The figure shows a labeled line chart with one visible annotation. Compare the callout mark with "Hari". What is the nonnegative difference between their values?

Answer: 7

Combo Mark: Conditioned Line Extremum



The figure shows stacked bars for "Nimbus" with an overlaid line for "Birch"; each category displays the relevant printed values. First keep categories with "Nimbus" greater than 55; which remaining category has the maximum "Birch" value?

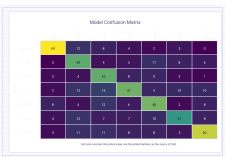
Answer: QV2

Uncertainty Band: Band Overlap

The image shows a line chart with two labeled series, each shown with a central line and a shaded upper-to-lower uncertainty band. At how many x-axis labels do the shaded bands for "Direct" and "Antimony" overlap?

Answer: 1

Matrix: Off-Diagonal Confusion



The figure shows a labeled confusion matrix with printed integer counts in each active cell. In the row for actual class "Tartus", exclude the matching diagonal cell. What predicted column label has the largest count?

Answer: Namanga

Figure 11: Representative chart tasks.

Sokoban: Box Goal Status

The figure shows a Sokoban board with walls, a player, colored boxes, and colored goal dots; a box on its matching goal has the goal dot drawn on top of the box. How many boxes are not on their matching colored goal dots?

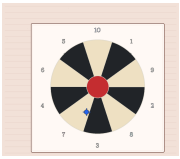
Answer: 3

Checkers: Max Capture Chain Length

Red moves on the shown 8 by 8 checkers board. A jump captures an adjacent opponent and lands on the empty square beyond. The marked piece is a king, so it may jump diagonally in either direction and continue while another capture is available. What is its maximum capture-chain length?

Answer: 2

Darts: Dart Score



The image shows a simplified labeled dartboard with visible dart markers. Using the simplified dartboard scoring rule, what is the dart's score? A dart in a numbered sector scores that number; a dart in the center bullseye scores 50.

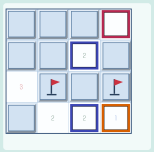
Answer: 7

Pac-Man: Next Item

This scene shows a Pac-Man-style maze with a visible Pac-Man marker, ghosts, pellets, bonus items, and a highlighted route. Which labeled bonus item is encountered first along the highlighted route?

Answer: F

Minesweeper: Forced Cell



In the Minesweeper grid, each number gives the mine count among its eight neighbors and each flag is a known mine. If a clue already has enough flags, its other hidden neighbors are safe; if every remaining hidden neighbor is needed, they are mines. From the outlined clue cells, how many hidden cells must be safe?

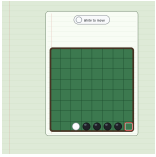
Answer: 5

Chess: Target Square Attacker

The scene shows a partial chess board with white and black pieces. How many black pieces currently attack the red outlined target square?

Answer: 0

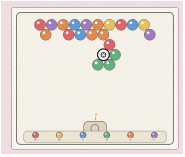
Reversi: Marked Move Flip



White moves on the shown 8 by 8 Reversi board. A legal move brackets opponent discs along one or more straight lines, and all bracketed discs flip. How many discs flip after playing at the red marked square?

Answer: 4

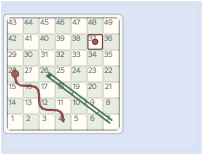
Bubble Shooter: Pop Color



A shot bubble attaches at the marked target. A connected same-color group pops when the placement makes its size at least three. Which option labels the bubble color that would pop a group there?

Answer: D

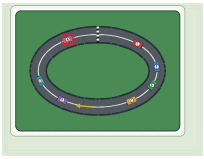
Snakes and Ladders: Remaining Distance to Finish



The visual shows a numbered Snakes and Ladders board with one token and visible snakes and ladders. How far is the token from the final square by square number?

Answer: 12

Racing Track: Ahead Object



The scene shows a top-down loop racing track with a checkered finish line, a direction arrow, a marked car, and other labeled cars. Follow the arrow direction from the marked car to the finish line. Do not count the marked car. Count only the cars ahead of the marked car before the finish line.

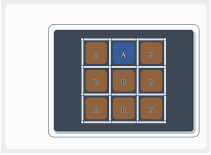
Answer: 0

Circular Chess: Target Cell Reacher

On this four-ring, sixteen-sector circular chess board, sectors wrap but rings do not. The red cell is the target. Rooks move along or across rings, bishops diagonally, queens as both, knights jump, and kings one step; sliding pieces stop at the first occupied cell. How many Black pieces can reach the target?

Answer: 3

Dots and Boxes: Owned Box

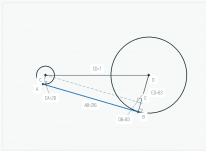


The scene shows a dots-and-boxes grid with some edges already drawn. How many boxes are marked with player B?

Answer: 8

Figure 12: Representative game tasks.

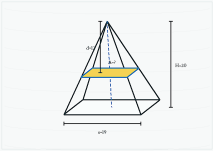
Circle Pair Tangents: Center Distance



This geometry diagram shows two separated non-concentric circles with centers C and D, a common tangent touching them at A and B, radii to the tangent points, and an auxiliary right triangle CED whose leg ED shows the radius difference. Find the center-to-center distance CD.

Answer: 225

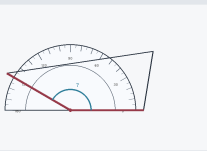
Solid Cross Section: Square Pyramid Parallel Slice Area



The labeled diagram shows a square pyramid cut by a plane parallel to its base, with the cross-section marked and dimensions labeled. The marked slice is parallel to the base. Use similar solids with scale d/H to find the value of A for the cross-section.

Answer: 130

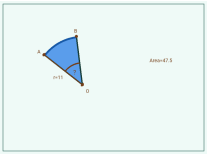
Measuring Tools: Protractor Angle



The figure shows a geometric figure with a protractor placed at the marked angle. Read the protractor scale at the marked angle. What is the angle measure in degrees?

Answer: 150

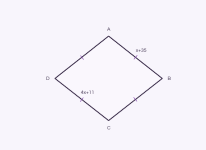
Sector: Central Angle from Sector Measure



The sector figure shows a circular sector with labeled points. Compute the measure of angle AOB. Use the labeled radius and shaded sector area.

Answer: 45

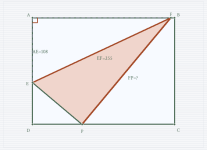
Special Quadrilateral: Segment Length



The figure shows a quadrilateral with construction marks and measurement labels. Solve the visible rhombus side expressions. What is the length of "segment CD"?

Answer: 43

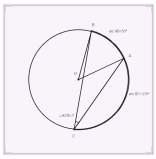
Paper Fold: Folded Segment Length



This folded-paper diagram shows a folded paper corner with dashed original edges, a crease, a folded flap, and side-length labels. The crease EF folds A onto P. What is the value of the marked segment FP?

Answer: 231

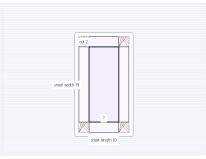
Circle Theorem: Inscribed Angle from Arc



The geometry diagram shows a circle with labeled points and shown measurements. Find the missing inscribed angle ACB.

Answer: 25

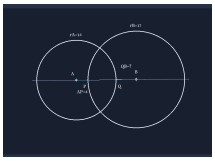
Rectangular Solid: Open Box Net Dimension



The diagram shows an open-box net with sheet dimensions, equal corner cutouts, and one missing base dimension. Using the sheet dimensions and cut size, find the marked base dimension.

Answer: 6

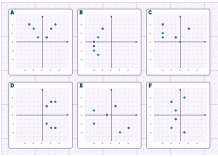
Circle Centerline Overlap: Segment Length



The diagram shows two overlapping circles whose centers lie on one line. Boundary labels mark where circles cross the shared centerline; radius labels and centerline segment labels are visible. Determine QB.

Answer: 7

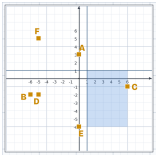
Coordinate Panels: Point Set Transform Match



The visual shows labeled coordinate panels, each containing a source point set and a candidate point set. Which labeled panel contains a source point set and its y-axis reflection?

Answer: A

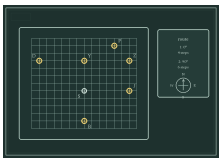
Coordinate Plane: Locus Point



The diagram shows a coordinate grid with one shaded locus region and lettered candidate points. Choose the lettered point inside the shaded region.

Answer: C

Bearing Route: Endpoint Position

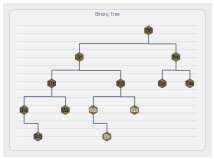


The figure shows a compass-bearing route instruction panel, a graph-paper candidate grid where each square is one step, the start point marked S, and labeled candidate endpoint points, with bearings measured clockwise from north. Follow the two bearing instructions from S. Which candidate endpoint label is reached?

Answer: Z

Figure 13: Representative geometry tasks.

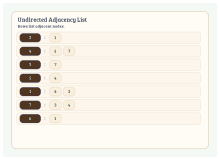
Binary Tree: Local Relative Node



The image shows a labeled binary tree with the root at the top and left/right children shown by position. Which node shares the same parent as "Max"?

Answer: Dak

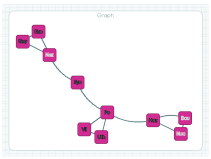
Adjacency: Undirected Component



You are shown an undirected graph as an adjacency list. Count the connected components in the undirected graph.

Answer: 2

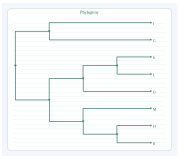
Node Link: Bridge



The figure shows a labeled undirected graph. What is the number of bridges in the graph?

Answer: 3

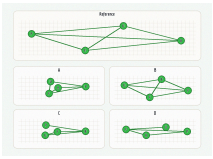
Phylogeny Tree: MRCA Clade Membership



The figure shows a rooted phylogeny cladogram with labeled taxa. Count the terminal taxa in the MRCA clade for "B" and "L".

Answer: 6

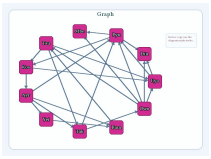
Graph Options: Same Structure



You are shown a Reference labeled graph above four labeled graph options. Which option has the same labeled graph structure as the Reference, ignoring layout?

Answer: B

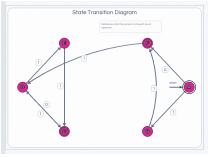
Node Link: Reachable Count After Edge Edit



The diagram shows a labeled directed graph. Remove the arrow from node "Uyo" to node "Pyu". Following arrow directions from "Uyo", how many nodes would be reachable?

Answer: 2

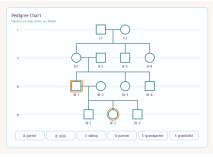
Automaton: State After Input



The figure shows a deterministic state-transition diagram with a start arrow, double-ring accepting states, and transition labels. Read input "1110" from the start state. What state label is reached after 2 symbols?

Answer: F

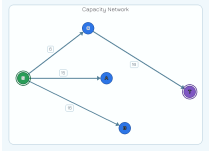
Pedigree Chart: Relationship



The diagram shows a labeled pedigree chart with family connections and answer options. Read the pedigree chart. Which option is III-1 relative to IV-2?

Answer: A

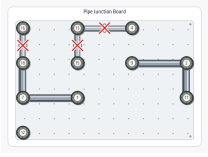
Flow Network: Max Flow



You are shown a directed capacity network with source S and sink T. Find the max-flow value from source S to sink T in this capacity network.

Answer: 6

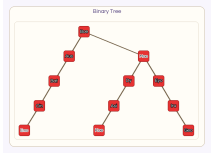
Pipe Network: Pipe Reachable Junction



The visual shows a labeled pipe-junction network with open pipes and blocked pipes (blocked pipes are marked with a red X). From 2, how many junctions are in the same open-pipe connected region, including the start junction?

Answer: 3

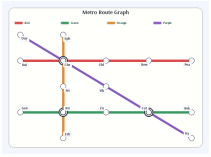
Binary Tree: Lowest Common Ancestor



This diagram shows a labeled binary tree with the root at the top and left/right children shown by position. Trace upward from "Eyo" and "Koe". What is the label of their lowest common ancestor?

Answer: Moe

Metro: Shortest Path Length



The figure shows a labeled metro route map with colored routes and stations. What is the shortest-path length in route segments from station "Pea" to station "Bek"?

Answer: 6

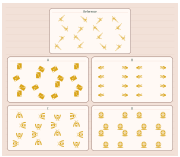
Figure 14: Representative graph tasks.

Named Path: Immediate Successor

The picture shows a START-to-END icon path with six option icons labeled A-F and other icons. Which labeled icon comes immediately after the first "flag" icon along the path from START to END?

Answer: F

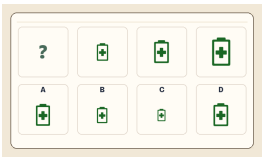
Wallpaper Panels: Same Pattern as Reference



The image shows one Reference wallpaper panel above four labeled wallpaper panels filled with repeated icon motifs. Select the labeled panel whose wallpaper pattern matches the Reference panel.

Answer: C

Sequence Strip: Size Progression Completion



The image shows a top row of four boxed sequence cells with one question-mark box and a bottom row of four labeled option boxes. Which labeled option should replace the question mark in the size sequence?

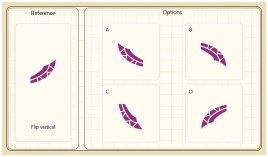
Answer: C

Icon Cutout: Partial Match

The image shows a partial icon fragment beside six labeled full-icon options. Which labeled full icon does the partial fragment come from?

Answer: B

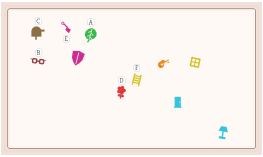
Single Transform Options: Inverse Geometric Transform Source



The scene shows a transformed Reference icon with a transform cue and four labeled source-option icons. Select the option that would become the Reference icon after a vertical flip.

Answer: B

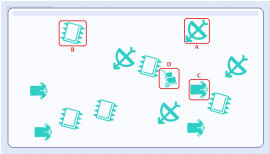
Named Field: Reference Distance Rank



The picture shows a panel with one reference icon, six option icons labeled A-F, and other icons. Which labeled icon is second closest to the magenta shield icon?

Answer: E

Icon Field: Frequency Extreme Type



The picture shows one panel of assorted icons. Which marked icon type appears most often in the image?

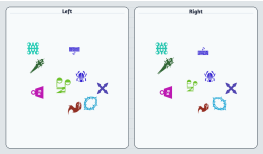
Answer: B

Icon Grid: Distinct Color

The scene shows a visible grid of icon cells. How many unique icon colors appear in the grid?

Answer: 4

Paired Canvas: Rotation Change



The picture shows two icon panels labeled Left and Right with corresponding icons at matching positions. How many Right-panel icons changed rotation compared with the corresponding icon in the Left panel?

Answer: 4

Venn Field: Same Region as Reference

The scene shows icons and two overlapping marked circles. By icon center, the possible regions are left-only, right-only, both-circle overlap, and outside both circles. Find the number of "mushroom" icons in the marked reference icon's region.

Answer: 3

Overlap Grid: Occlusion Order

The image shows a Reference cell beside a labeled Scene grid. How many labeled Scene cells match the Reference front-to-back order? Front-to-back order means which icon is on top.

Answer: 3

Named Grid: Scoped Attribute

The image shows a numbered grid with one icon in each cell. How many "trash can" icons are in row 4?

Answer: 1

Figure 15: Representative icon tasks.

Library: Filtered Book in Section

The scene shows a library room containing 5 labeled book sections. How many upright books are in the Music section?

Answer: 3

Indoor Room: Swapped Tile Pair

The scene shows a bedroom with small objects. Find the pair of numbered cells that were swapped in the room image.

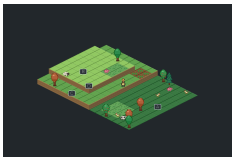
Answer: A

RPG Dungeon: Safe Reachable Chest

The scene shows a top-down RPG-style dungeon map with walkable floor, blocked passages, treasure chests, and monsters inside some chambers. Starting from the player, how many treasure chests can be reached without crossing boulders, excluding any chest in a chamber that contains a monster?

Answer: 1

Isometric Farmstead: Farmer Same Level Tile



This illustration shows a farmstead drawn on isometric tiles, with a farmer and lettered candidate ground tiles. Select the lettered farm tile whose ground level matches the farmer's tile.

Answer: C

Park Playground: Missing Patch

The picture shows a synthetic park playground. Select the option that fills the missing patch exactly.

Answer: D

Pixel Village: Rotated Tile

The canvas contains a top-down pixel village scene. Which letter marks the tile with the incorrect orientation?

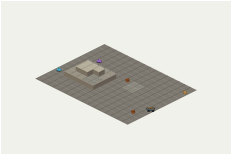
Answer: C

RPG Tactical Map: Water-Barrier Unreachable Tile

Water tiles form a barrier and cannot be crossed; every other tile can. The blue unit moves only up, down, left, or right. Which candidate letter marks a tile the blue unit cannot reach?

Answer: D

Isometric Quarry: Highest Terrain Tile



The scene shows a pixel-art isometric quarry with a clean highest terrace and lower surrounding rock terrain. Count only the visible top-surface tiles on the highest elevated quarry layer.

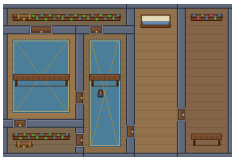
Answer: 5

Environment: Rotated Tile

The image shows a meadow river setting with illustrated foreground objects and visible environmental features. One of the lettered tiles is rotated relative to the rest of the environment image. Which tile is it?

Answer: D

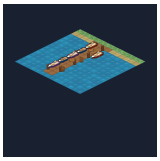
RPG House: Adjacent Room



The canvas contains a top-down RPG house map with enclosed rooms, walls, and doorways. How many rooms are immediately connected to the room with the player by a doorway?

Answer: 4

Isometric Harbor: Boat Side



The picture shows an isometric dockyard with shoreline tiles, water tiles, a wooden pier, dock objects, and boats tied along the dock. Count only the boats on the image-left side of the wooden main dock.

Answer: 5

Construction Site: Equipment Zone

The canvas contains a construction site with 5 workers, labeled work zones, materials, and equipment. How many construction vehicles are in the Roadwork Zone? If none are there, answer 0.

Answer: 0

Figure 16: Representative illustration tasks.

Web Action: Action Target

The browser page includes a guide-code table and labeled candidate controls. Click the control on the item with category "Home", status "Clearance", and guide code "M2". Which candidate label marks that control?

Answer: S

Mixed Infographic Page: Field-Total Comparison

The image shows a dense mixed infographic page with titled modules, item labels, field labels, printed values, and native context text. Between "Music" and "Logistics", which module has the greater sum of "Score" values?

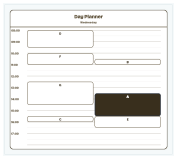
Answer: Music

Infographic: Section Icon Total Difference

The image shows a multi-section infographic with labeled metric cards and printed values. Add the credit card-icon cards in section "Education" and in section "Beverage". What is the nonnegative difference?

Answer: 70

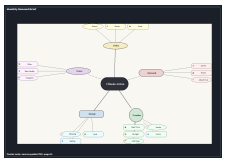
Schedule: Longer Than Reference



This schedule shows one single-day schedule with time labels and scheduled event blocks. How many other scheduled events are longer than the highlighted reference event?

Answer: 0

Concept Map: Branch Child



The concept map shows a central topic, labeled branches, visible child-item nodes, and connector lines. How many child items are listed under the "Weather" branch?

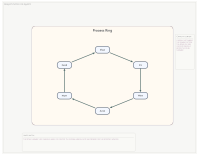
Answer: 5

Hero Callout Infographic: Callout Composite Metric Extremum

The figure shows titled callout cards with field labels, visible values, badges, and connector lines. Order the callouts by "Score" plus "Count", highest to lowest. Which callout title is first?

Answer: Automotive

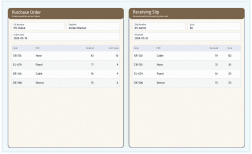
Cycle: Offset Stage



This diagram shows a directed cycle with labeled stages connected by arrows. Use the arrows and count 2 steps after "Gord". What stage do you reach?

Answer: Iris

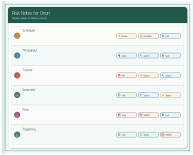
Paired Forms: Total Amount Delta



The purchase order and receiving slip show matching item codes, quantities, and purchase-order unit values. For each item with differing quantities, multiply the absolute quantity difference by its unit value and sum the products. What is the total amount delta?

Answer: 80

Instruction Panel: Step for Control Pair



The visual shows an instruction panel with numbered steps and visible control chips on each step. Find the step with control labels "Search" and "Share". What number is on that step?

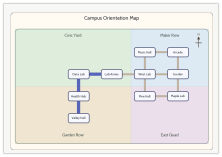
Answer: 2

Step List: Relative Offset Step

This page shows a numbered workflow step list where each step card has labeled Title and Detail fields plus other small fields. What step title is 3 steps after the step titled "Payload"?

Answer: Drift

Map: Landmark After Route Step



The figure shows a printed campus layout with labeled landmarks, named zones, visible walking paths, and a highlighted orange route. Trace the highlighted path from "Valley Hall" to "Lab Annex". Which label is at the 3rd reached landmark?

Answer: Lab Annex

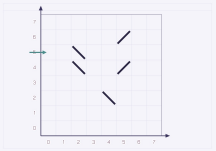
Profile Card Grid: Highest-Value Profile

This page shows a grid of profile cards, each with a visible profile name and labeled fields. Which visible profile ranks highest by Hours?

Answer: Patsyann

Figure 17: Representative page tasks.

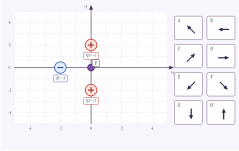
Ray Optics: Ray Bounce



This figure shows a reflected ray setup on a square grid. What is the total number of mirror bounces in the implied ray path?

Answer: 4

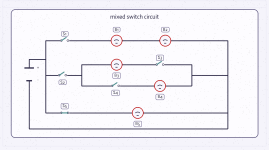
Electrostatic Field: Field Direction Choice



The diagram shows an electrostatic coordinate grid with fixed point charges, a marked point or labeled candidate points, and visible option arrows or distance labels. Which option A-H points in the direction of the electric field at P?

Answer: B

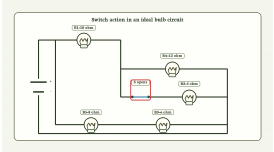
Switch Circuit: Lit Bulb



The circuit diagram shows a mixed branch battery circuit with five labeled bulbs and visibly open or closed switches. Given the switch states, how many bulbs will be on?

Answer: 1

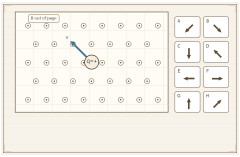
Circuit State Change: Bulb Brightness Change



The image shows a visible ideal-battery circuit with five labeled bulbs, resistance labels, and one red-boxed switch action cue. Using the bulb resistances and topology, which bulb becomes brighter after the switch action?

Answer: B4

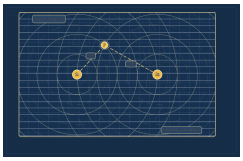
Magnetic Force: Force Direction Choice



The image shows a magnetic-field panel with field-direction symbols, a charged particle, a velocity arrow, and eight labeled candidate force arrows. Which option A-H points in the direction of the force on the charged particle?

Answer: H

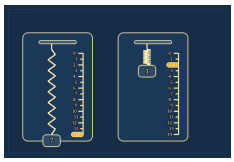
Wave Interference: Path Difference



The diagram shows two labeled wave sources and their circular interference wavefronts. Using the visible wavefront spacing and labels, determine the absolute path difference between S1P and S2P in half-wavelength steps.

Answer: 2

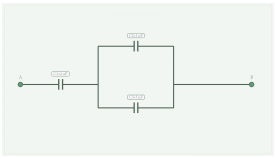
Spring: Spring Extension Difference



The figure shows a pair of identical springs with weight blocks and ruler markers. Using the two ruler markers, what integer difference in extension do the springs show?

Answer: 12

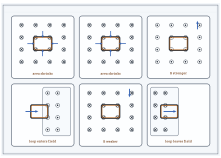
Circuit Equivalent: Total Capacitance



The visual shows one capacitor circuit between terminals A and B with at least one labeled parallel-plate capacitor in series with one or two labeled parallel capacitor blocks. Using the shown capacitor values, what equivalent capacitance is seen between terminals A and B?

Answer: 2

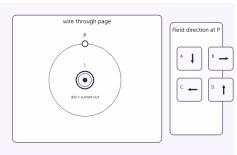
Electromagnetic Induction: Induced Current Direction



The induction diagram shows six mini-panels showing conducting loops, into-page or out-of-page magnetic-field symbols, and visible cues for changing or unchanged magnetic flux. How many panels have a clockwise induced current in the loop?

Answer: 6

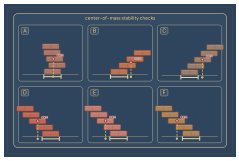
Wire Magnetism: Wire Field Direction Choice



The diagram shows a current-carrying wire through the page, point P, and labeled arrow options. Use the current cue to determine the circular magnetic-field direction at P. Which arrow option matches that direction?

Answer: C

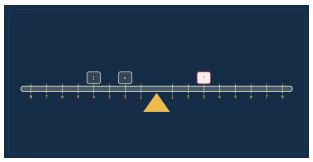
Stack Stability: Stability Status



The image shows six labeled stacks of equal-size brick blocks. Which labeled stack will stay upright because its center-of-mass projection falls within its support base?

Answer: A

Lever: Missing Weight Balance

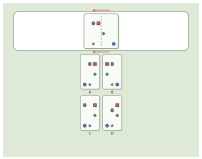


The figure shows a lever with a fulcrum, distance marks, and weight blocks. What weight should replace the block marked with a question mark so the lever balances?

Answer: 4

Figure 18: Representative physics tasks.

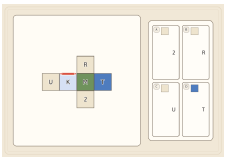
Sheet Transform: Fold Projection Result



The visual shows an outline-style paper-folding puzzle with a fold diagram above labeled result options. After the shown fold is made, which labeled option shows the visible mark pattern?

Answer: A

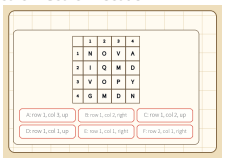
Cube Net: Marked Edge Neighbor Face



The puzzle diagram shows a labeled cube net with one red marked edge and labeled face options. Which labeled option touches the red marked edge on the folded cube?

Answer: B

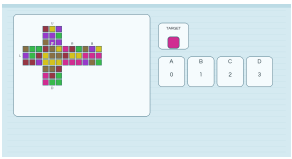
Word Search: Search Location



The visual shows a row-and-column labeled word-search grid. Which labeled option below the grid matches the placement of "NOVA" in the word-search grid?

Answer: E

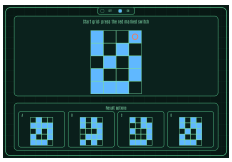
Rubik's Cube Net: Post-Move Face Color Count



The scene shows a Rubik's Cube net with face labels, a target-color swatch, and four numbered options. A prime mark denotes a counterclockwise turn as viewed from outside the affected face. After the sequence R' B' L, which option gives the number of target-color stickers on the upper face?

Answer: B

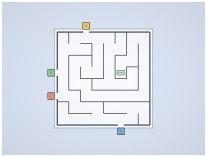
Toggle Grid: Toggle Result



The scene shows a toggle grid with switch cells, binary cell states, and visual answer options. Use the start grid and the red marked switch. After that one press toggles its cell and edge-neighbor cells, which labeled option matches the final state?

Answer: D

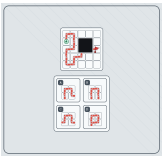
Maze: Nearest Exit



The image shows an orthogonal wall maze with a START cell and labeled exits on the outer boundary. Which labeled exit is nearest to START by corridor path length?

Answer: A

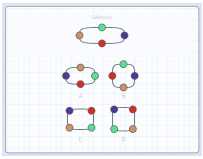
Pipe Flow: Repair Tile



The pipe puzzle shows an industrial conduit tile grid with a green start marker and red finish flag. Find the single option whose displayed openings make a continuous path from the green start marker to the red finish flag through the black 2x2 gap. Do not rotate or flip options.

Answer: A

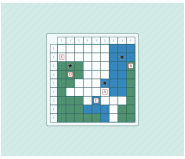
Cyclic Order: Equivalent Arrangement



The image shows a reference token loop above six labeled option loops. Use the token colors when comparing cyclic order. Exactly one option loop is equivalent to the reference loop. Rotations count as equivalent; reversed or reflected order does not.

Answer: D

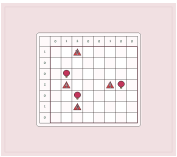
Star Battle: Valid Cell Anywhere



The visual shows a partially filled Star Battle grid with colored regions, visible stars, and labeled candidate cells. Use these Star Battle rules: each row, column, and colored region must contain exactly one star; stars cannot touch, even diagonally. Choose the labeled cell where another star can be placed.

Answer: C

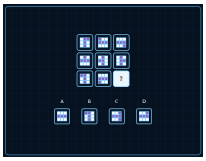
Tents: Violating Tent



The figure shows a Tents puzzle grid with row and column clues, trees, visible tents, and labeled cells or tents. Inspect the labeled tents. Which label marks the tent that has no orthogonally adjacent tree?

Answer: D

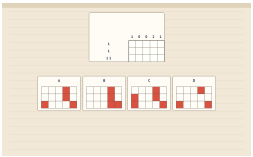
Raven Matrix: Spatial Transform



The image shows a 3 by 3 Raven-style matrix puzzle with four labeled image options below it. Select the option whose mini-grid pattern fits the spatial arrangement rule.

Answer: D

Nonogram: Candidate Solution

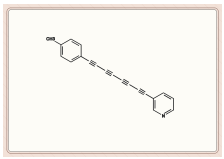


This nonogram shows a clue grid with labeled filled-grid options. Use both clue rails to identify the correct candidate grid.

Answer: A

Figure 19: Representative puzzle tasks.

Organic Structure: Bond Order



The visual panel contains an organic skeletal structure with line-angle bonds, rings, and occasional atom or substituent labels. Count only the visible triple bonds.

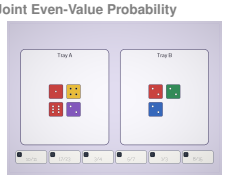
Answer: 4

Life Automaton: One Step Cell State

In the START grid, dark cells are alive. An alive cell survives with two or three live neighbors; an empty cell becomes alive with exactly three; every other cell is empty next. After one update, how many cells are empty?

Answer: 9

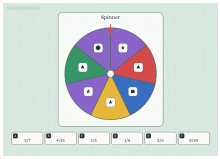
Dice: Joint Even-Value Probability



The trays show each die's top value. One die is selected uniformly and independently from each tray. Which A-F option gives the probability that both selected dice show even values?

Answer: C

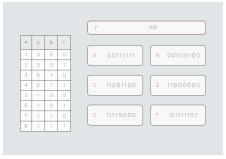
Spinner: Color-and-Shape Probability



Each spinner sector is equally likely and has a color and shape. Which A-F fraction option gives the probability of landing on a purple sector marked with a star?

Answer: A

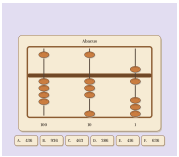
Truth Table: Truth Pattern



A truth table shows an input table, a target expression P, and six labeled output-pattern options. Values use 1 for true and 0 for false. The operators are ! for NOT, & for AND, | for OR, and ^ for XOR. Reading rows from top to bottom, which option matches the truth pattern for P?

Answer: A

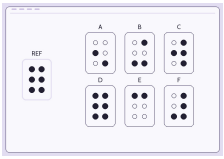
Abacus: Displayed Value



On the three-column abacus, beads touching the center bar are active. An active upper bead is worth 5 and each active lower bead 1. Read the 100, 10, and 1 columns from left to right. Which option gives the displayed number?

Answer: A

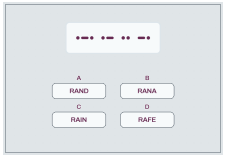
Braille Cell: Matching Pattern



This panel shows a Braille reference cell and six labeled Braille option cells. Which option letter matches the reference pattern?

Answer: D

Morse Code: Morse Word Read



A visual Morse-code panel presents a Morse-code word above four labeled word options. Which labeled option matches the Morse code?

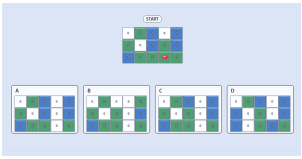
Answer: C

Music Staff: Scale Degree Function

The visual shows a sheet-music staff with marked notation items. Select the option naming the scale-degree function of note 3 in G major.

Answer: E

Agent Automaton: Future Grid



Starting at the arrow, apply three updates. Before moving, state 0 turns right and becomes 1; state 1 goes straight and becomes 2; state 2 turns left and becomes 0. The agent then moves one cell forward, wrapping at the grid edge. Which option shows the resulting grid?

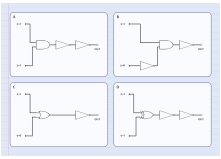
Answer: C

Turing Tape: Written-Symbol Count

Starting from the shown tape, head, and state, simulate six transitions. At each step, use the visible table to write a symbol, move left or right, and change state. How many tape cells contain symbol 0 afterward?

Answer: 5

Logic Gate Circuit: Output Value



A Boolean logic-gate panel shows a circuit with labeled inputs, standard gate symbols, wires, and final OUT nodes. Evaluate each circuit. Which option's final OUT value is 1?

Answer: A

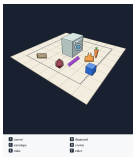
Figure 20: Representative symbolic tasks.

Object Cluster: Color-and-Type Exclusion

The visible scene contains many small 3D objects arranged on a plain surface. How many cyan objects are not cups?

Answer: 1

Object Scene: Nearest to Reference



The scene shows a washing machine as the reference and six smaller candidate objects. Which option identifies the candidate nearest the washing machine?

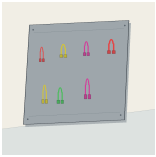
Answer: F

Object Cluster: Type-or-Color Count

The image shows many small 3D objects arranged on a plain surface. How many objects are either stools or blue?

Answer: 5

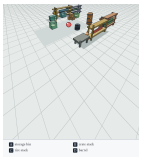
Surface Fixture: U-Bolt Count



The image shows a U-bolt plate with visible U-bolts. How many U-bolts are visible?

Answer: 7

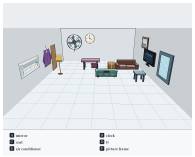
Warehouse: Nearest Candidate to Reference



The image shows a perspective warehouse aisle with shelf racks, warehouse equipment, one red sphere reference object, candidate warehouse objects, and a text option panel below the scene. Which option describes the warehouse object closest to the red sphere?

Answer: C

Room: Object on the Reference Wall



The picture shows a perspective indoor room with floor, walls, furniture, room objects, and wall-mounted objects. Which option describes the object that shares a wall with the fan?

Answer: B

Carousel: Ordered Adjacent Pair Count

The scene shows small 3D objects on two concentric elliptical belts. Following the order of the inner belt, count each occurrence of gloves immediately followed by pyramids. How many pairs are there?

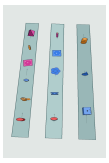
Answer: 0

Street: Intersection Nearest

The view shows a perspective street intersection or T intersection with roads, sidewalks, crosswalk markings, street context, candidate street objects, and a text option panel below the scene. Which option describes the object closest to where the roads meet?

Answer: B

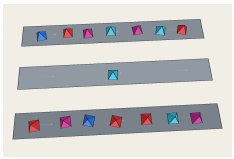
Conveyor: Ordered Adjacent Pair Count



The visual shows a three-lane conveyor with small objects on each lane. Following the order of the left belt, count each occurrence of flowers immediately followed by trophies. How many pairs are there?

Answer: 0

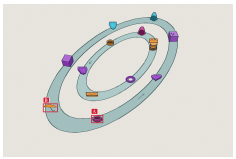
Conveyor: Objects on Selected Belt



The visual shows a three-lane conveyor with small objects on each lane. How many objects are on the MIDDLE belt?

Answer: 1

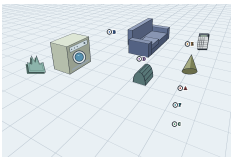
Carousel: Between Object Type Anchors



A 3D conveyor carousel is shown with objects on the inner and outer belts. Follow the OUTER belt direction from the marked bowl A to the marked umbrella B. How many objects are strictly between them?

Answer: 5

Object Scene: Marked Point Depth Extremum



The scene shows a perspective 3D platform scene with objects and lettered floor point markers. From this camera view, which lettered floor point is farthest away from the viewer?

Answer: B

Figure 21: Representative 3D tasks.