



ZEBRA-CoT: A Dataset for Interleaved Vision-Language Reasoning

Ang Li^{*▲} Charles L. Wang^{*▲} Kaiyu Yue^{*♥} Zikui Cai^{*♥} Ollie Liu^{*▲} Deqing Fu^{*▲}
 Peng Guo^{*♦} Wang Bill Zhu^{*●} Vatsal Sharan^{*} Robin Jia^{*} Willie Neiswanger^{*}
 Furong Huang[♥] Tom Goldstein[♥] Micah Goldblum^{*}

^{*}Columbia University [♥]University of Maryland [♦]University of Southern California

[△]New York University

^{*}Equal contribution

Abstract: Humans often use visual aids, for example diagrams or sketches, when solving complex problems. Training multimodal models to do the same, known as Visual Chain of Thought (visual CoT), is challenging due to: (1) poor off-the-shelf visual CoT performance, which hinders reinforcement learning, and (2) the lack of high-quality visual CoT training data. We introduce **ZEBRA-CoT**, a diverse large-scale dataset with 182,384 samples, containing logically coherent interleaved text-image reasoning traces. We focus on four categories of tasks where sketching or visual reasoning is especially natural, spanning *scientific questions* such as geometry, physics, and algorithms; *2D visual reasoning tasks* like visual search and jigsaw puzzles; *3D reasoning tasks* including 3D multi-hop inference, embodied and robot planning; *visual logic problems and strategic games* like chess. Fine-tuning the Anole-7B model on the ZEBRA-CoT training corpus results in an improvement of +12% in our test-set accuracy and yields up to +13% performance gain on standard VLM benchmark evaluations. Fine-tuning Bagel-7B yields a model that generates high-quality interleaved visual reasoning chains, underscoring ZEBRA-CoT’s effectiveness for developing multimodal reasoning abilities. We open-source our dataset and models to support development and evaluation of visual CoT.

Datasets: multimodal-reasoning-lab/Zebra-CoT

Anole-Zebra-CoT Model: multimodal-reasoning-lab/Anole-Zebra-CoT

Bagel-Zebra-CoT Model: multimodal-reasoning-lab/Bagel-Zebra-CoT

GitHub Repository: github.com/multimodal-reasoning-lab/Bagel-Zebra-CoT

1 Introduction

Human cognition naturally integrates multimodal thought processes when solving complex problems. For example, a high school student sketches diagrams to solve geometry or physics problems, an engineer creates diagrams to design and debug workflows, and a data scientist generates plots to better understand data. These visual aids are central to effective problem solving. While recent vision-language models (VLMs) have shown strong performance on multimodal tasks like visual question answering, their reasoning traces remain predominantly textual. Enabling models to explicitly reason in the visual space, Visual Chain of Thought (visual CoT), remains a fundamental open challenge. Unlocking visual CoT may improve reasoning performance in domains where visual intuition is relevant and may make the reasoning patterns expressed by models more interpretable to humans.

Recent advances in frontier multimodal models (Team et al., 2023; Hurst et al., 2024; Bai et al., 2025; OpenAI, 2025a; Team, 2024; Chern et al., 2024; Sun et al., 2024; Deng et al., 2025) have made visual CoT feasible primarily through agentic pipelines that leverage external tools (e.g., Python functions, or expert vision models) for visual programming (Surís et al., 2023), such as generating sketches for geometry, algorithms, and spatial reasoning tasks (Hu et al., 2024; OpenAI, 2025b), or bounding boxes for fine-grained visual tasks (Shao et al., 2024a; Wu and Xie, 2024; Zheng et al., 2025). An emerging possibility is innate visual reasoning where models directly generate explicit visual tokens during their thinking process (Li et al., 2025; Chern et al., 2025; Xu et al., 2025b). However, current VLMs with interleaved text and image generation capabilities (Team, 2024; Chern et al., 2024) either fail to generate useful visual aids for reasoning, or are not trained for such multimodal generation inherently during the reasoning process (Deng et al., 2025), making reinforcement learning approaches to reasoning infeasible. Li et al.

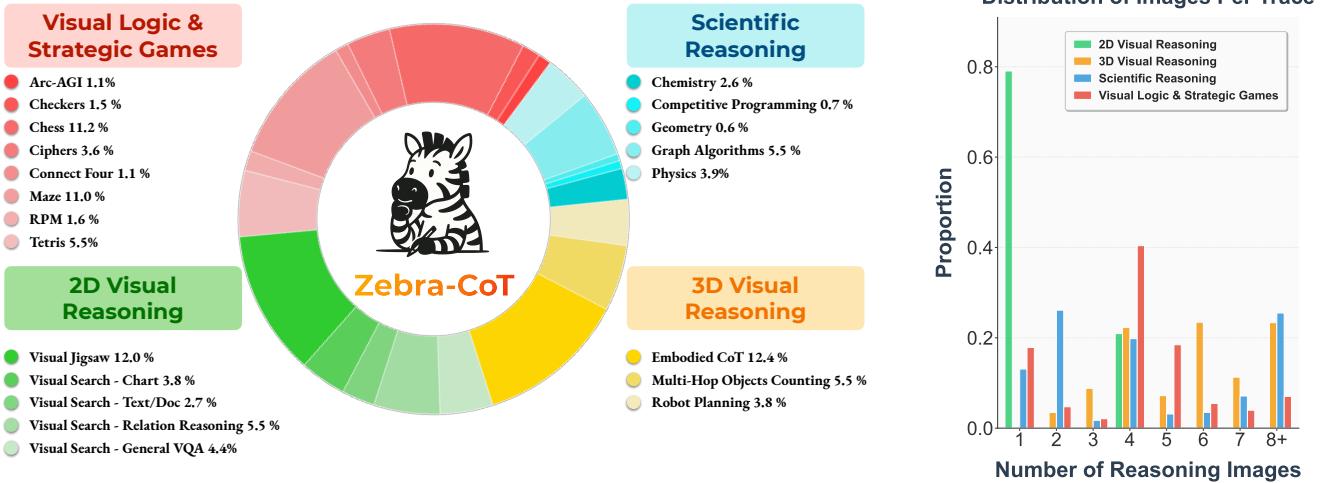


Figure 1 We curate a large-scale multimodal dataset by sourcing and cleaning raw traces from real-world domains, and generating synthetic examples using templated reasoning filled in by VLMs. ZEBRA-COT comprises 4 major categories and 18 subcategories, encompassing over **182K** instances in total. A detailed breakdown of the data statistics appears in [Table 3](#).

(2025) demonstrate visual CoT in synthetic mazes by training specialist models, but we remain far from foundation models capable of general high-quality visual CoT, largely due to the lack of large-scale diverse interleaved text and image reasoning training datasets.

To support the development of next generation vision language models that can explicitly reason with both text and visual modalities, we present **ZEBRA-COT**, a high quality dataset of interleaved text and image reasoning traces. Our dataset covers four main categories: scientific questions, 2D visual reasoning, 3D visual reasoning, and visual logic and strategic games, each containing multiple subdomains and task types as exemplified in [Figure 2](#). To the best of our knowledge, ZEBRA-COT is the first dataset to provide diverse and logically coherent multimodal reasoning traces across such a wide range of domains. Unlike prior large-scale interleaved datasets that are primarily composed of web-scraped image-text pairs with weak semantic alignment and no explicit reasoning structure (Li et al., 2024b; Awadalla et al., 2024; Zhu et al., 2023), ZEBRA-COT is carefully curated as a training resource in the spirit of high-quality text-based reasoning datasets. At the same time, compared to the only existing open-source interleaved text visual reasoning dataset we are aware of, VISUAL-COT (Shao et al., 2024a), which focuses on a single task of visual search, ZEBRA-COT introduces a much broader and more diverse set of tasks with richer reasoning trajectories. We provide a detailed comparison with other datasets below in [Table 1](#).

In total, ZEBRA-COT contains 182,384 samples. After fine-tuning ANOLE-7B (Chern et al., 2024) on our training set, we improved the accuracy on our in-distribution test set from 4.2% to 16.9%, delivering a **4 times** relative performance improvement and a **12%** gain in accuracy. When evaluating with benchmarks requiring visual reasoning, our anole model achieves an average of **4.9%** improvement across seven challenging datasets, with a maximum gain of **13.1%** on a visual logic benchmark, as shown in [Table 2](#). Furthermore, we fine-tune our dataset on BAGEL-7B (Deng et al., 2025), a high-quality multimodal model that cannot in its original form generate interleaved text and images. After fine-tuning, the model is able to inherently generate high-quality visual CoT during its own reasoning process, making it well-suited for future RL training, as shown in qualitative examples in [Section 5](#). We release the weights of both models to facilitate further research.

2 Related Work

Visual chain of thought. The community has predominantly been tackling visual CoT by using visual programming to generate images (Surís et al., 2023; Zhang et al., 2023; Mitra et al., 2024; Yang* et al., 2023; Wu and Xie, 2024; Hu et al., 2024; Menon et al., 2024; OpenAI, 2025b; Zheng et al., 2025). In particular, VISUAL SKETCHPAD (Hu et al., 2024) presents the most versatile open-source visual reasoning agents among existing works, handling a wide range of tasks. Another line of work explores model-generated images: for example, Rose et al. (2023) uses a diffusion model to bridge gaps in storytelling, and Chern et al. (2025) generates intermediate images to improve

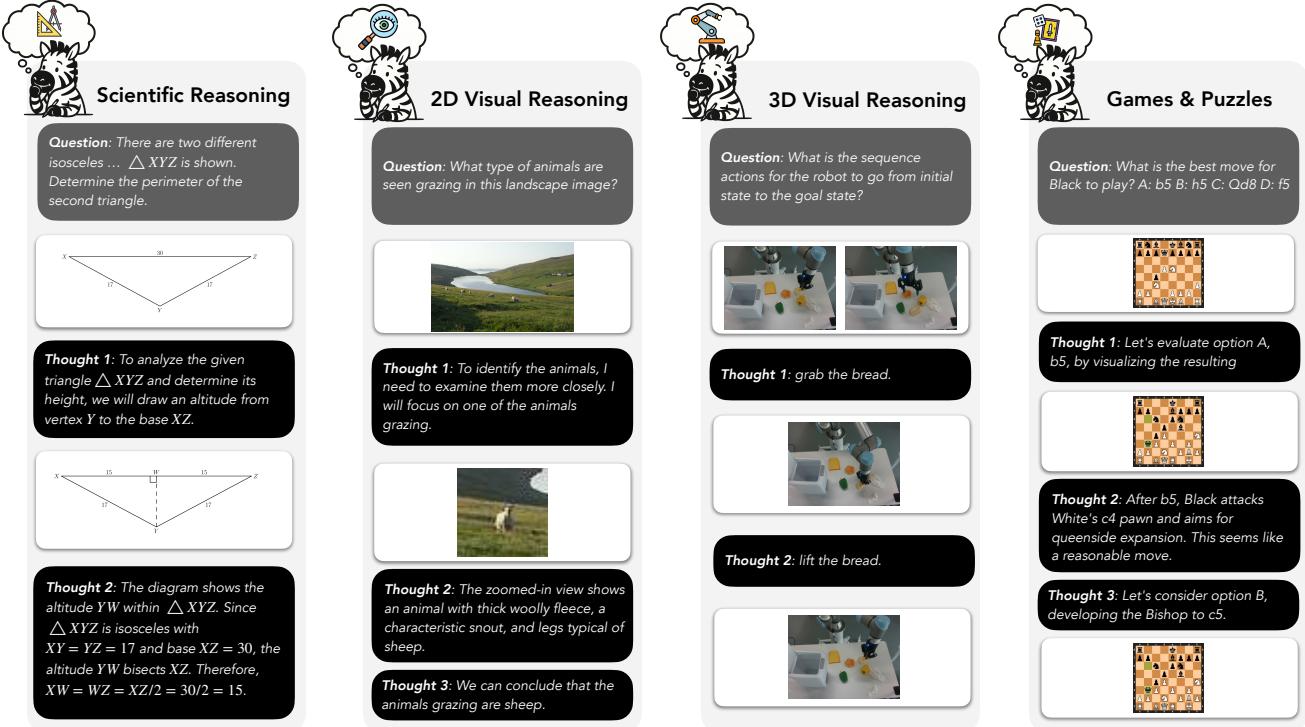


Figure 2 Visual CoT helps answer complex visual reasoning questions, as illustrated by examples from ZEBRA-CoT.

Dataset	Primary Task	Modality	Limitations
GQA	Compositional visual QA	Image, Text	No visual CoT
ScienceQA	Multimodal science QA	Image/Diagram, Text	No visual CoT
MM-PhyQA	Physics Visual CoT	Image, Text	Physics data only, not open sourced
Visual CoT	Visual-search QA with bbox CoT	Image, Text	Limited to visual search tasks
CoT VLA	Robotics Visual CoT	Image, Action	No text reasoning
R1-Onevision	A SFT and RL multimodal reasoning training dataset	Image, Text	No visual CoT
OmniCorpus	10 B-level interleaved corpus	Image, Text	Noisy pretraining data
MINT-1T	1 T-token web-scale interleaved data	Image, Text	Noisy pretraining data
ZEBRA-CoT	Diverse and high quality Visual CoT	Image, Text	Broad task coverage and CoT with explicit visual aids

Table 1 ZEBRA-CoT introduces a broader set of high quality visual CoT traces compared with prior datasets and pipelines.

image generation tasks; Zhao et al. (2025) generates intermediate images as subgoal predictions and derives actions based on them for robotic planning; Li et al. (2025) and Xu et al. (2025b) explore spatial reasoning tasks like mazes by visualizing each temporal step. However, these model-generated image approaches are mostly specialists, and developments are still primitive compared to visual programming methods that leverage external tools.

Visual reasoning datasets. Many multimodal visual reasoning datasets have been proposed (Lu et al., 2022; Wang et al., 2024c; Mu et al., 2023; Xu et al., 2024; Guo et al., 2024; Sun et al., 2025; Yang et al., 2025; Johnson et al., 2017; Zellers et al., 2019), although most focus on multi-modality only in the input question, leaving the reasoning traces purely textual. Among them, Shao et al. (2024a) stands out as the only open-source dataset featuring interleaved text and image reasoning. Anand et al. (2024) on the other hand, introduces a paradigm for incorporating images into the reasoning process for physics problems, though the dataset is not publicly available. Several vision-centric benchmarks (Fu et al., 2024b; Hao et al., 2025a) present diverse and challenging tasks, but they lack annotated reasoning traces.

Interleaved text and image datasets. Large-scale corpora with interleaved text and images have become essential for pretraining VLMs with reasoning capabilities (Alayrac et al., 2022; Chen and Wang, 2022; Sun et al., 2024; Wang et al., 2024b; Hurst et al., 2024; Li et al., 2024a; Bai et al., 2025; Team et al., 2025). However, in most existing interleaved text and image datasets Zhu et al. (2023); Laurençon et al. (2023), images are primarily used for

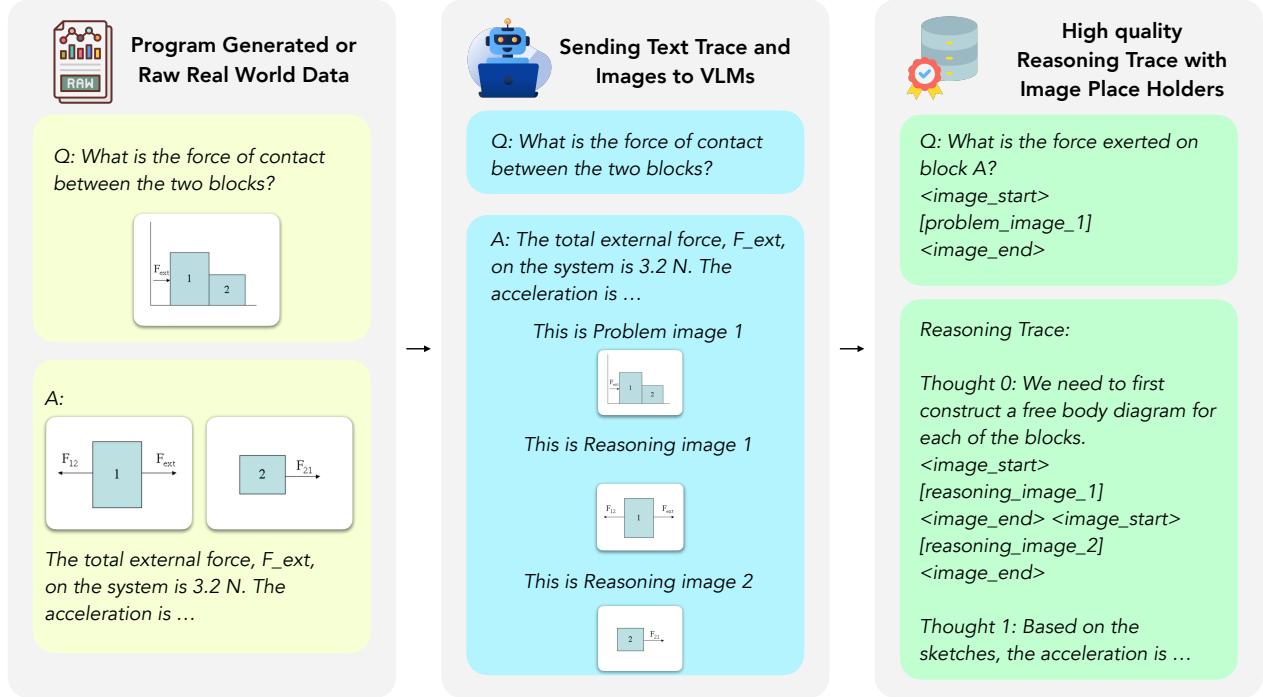


Figure 3 An overview of our data curation pipeline.

recognition, captioning, or as supplementary context in text-based reasoning, rather than serving as explicit visual aids that contribute meaningfully to the reasoning process. While Awadalla et al. (2024) include some scientific content from arXiv where images may aid reasoning, both the text traces and visual content are often noisy and not well-suited for post-training or fine-grained reasoning tasks. Instead, our ZEBRA-COT introduces a broader and higher-quality set of visual CoT examples, enabling effective training for visual reasoning.

3 Curating an Interleaved Text and Image Reasoning Dataset

Existing interleaved text and image datasets lack strong logical coherence between the two modalities. To train models capable of generating useful and accurate multimodal rationales, it is essential that the connection between text and images is both meaningful and instructive, especially why explicit visual thoughts are needed and how generating visual tokens can contribute to problem-solving. Moreover, such datasets should contain problem instances where visual aids are useful and visual tokens can contribute to problem-solving.

To address these requirements, we first source a diverse range of question types and domains. For real world data, we source high-quality problems from online resources, such as math, physics, coding, and chess competition datasets. We then extract and clean available raw reasoning traces containing text and images. However, even from high quality sources, traces can still lack clear logical connections between modalities, as well as clear references to the images for automatic parsing into interleaved text and image data ready for training. For example, most geometry data uses reference labels such as “Figure x”, which makes it hard to find the mapping between the actual image and the text reference. For synthetic data, we create our own examples by generating images or utilizing real images from online sources, then crafting corresponding reasoning templates. This procedure raises a clear issue, namely that we lack diversity and expressiveness of textual reasoning in templated data. For instance, in visual search tasks, it is crucial to elucidate the rationale behind drawing specific bounding boxes, and for chess, generating reflections and descriptions of move visualizations is key.

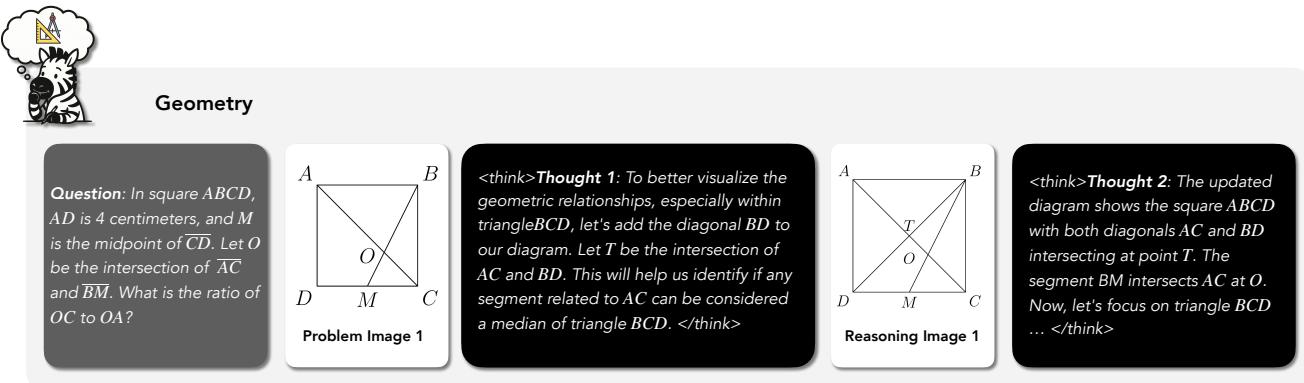
We address both of these issues using VLMs (Gemini-2.5 and GPT-4.1) to fill in the template placeholders or enhance the reasoning traces and complete the textual reasoning narrative. We feed in both images and raw text reasoning traces to the language model and ask the language model to output pure text traces with image placeholders. We further filter out invalid cases such as multiple image placeholders referring to the same image and unreferenced image placeholders to make sure that the data can be automatically parsed into a training dataset.

Existing multimodal rationale datasets are also limited in their breadth. The only available datasets focus on either visual search (Wu and Xie, 2024; Shao et al., 2024a) or spatial reasoning like maze navigation (Li et al., 2025). Such limited datasets are unlikely to enable training visual reasoning models that can generalize across domains more broadly. Visual Sketchpad (Hu et al., 2024) offers a diverse range of VLM agents to tackle a wider variety of questions. Though Sketchpad offers a powerful and significant contribution for generating visual aids, the pipeline is not designed for collecting post-training datasets. First, the reasoning traces generated by agentic pipelines often involve tool call errors and debug information, which degrade their quality. Second, the scalability and diversity of the dataset are fundamentally constrained by the limited number of agent tool designs and the high cost, as each reasoning trace may require many API calls. To tackle those issues, we curate in total over **182K** high-quality interleaved text and visual reasoning traces, spanning four major categories, including scientific reasoning, 2D visual reasoning, 3D visual reasoning, and visual logic and strategic games. We provide the details in the section below and example traces from our dataset.

4 Data Composition

4.1 Scientific Questions

Geometry. Geometric understanding is a core ability for multimodal models to ground reasoning over complicated mathematical tasks. Many datasets have been proposed to evaluate mathematics capabilities, including geometry. The MATH dataset (Hendrycks et al., 2021) is widely used for evaluating the mathematical performance of LLMs. Although the MATH dataset includes numerous geometry competition problems, their geometric elements are provided as plotting code rather than rendered images (see Figure 5). In ZEBRA-COT, we convert every piece of plotting code into figure renderings, producing both the problem diagram and its solution illustration to serve as an explicit visual reasoning chain for model training.



In total, we collect 1,061 samples from the MATH dataset’s train split. Our data provides only rendered images for both the problem and solution reasoning chains, with no plotting code included. Solving these problems requires generating images to assist. The problems are not restricted to the geometry subcategory but also include some problems from counting and probability, pre-algebra, pre-calculus, etc.

Physics. A variety of physics problems benefit from sketches, such as free body diagrams for force analysis, motion diagrams for kinematics, circuit diagrams for electricity, and ray diagrams in optics. We construct samples of classical mechanics problems programmatically. Problem instances are generated from parametric Python templates (e.g., Atwood machines, inclined planes, elastic collisions, pendulums), with physically plausible parameters sampled from predefined ranges. For each sample, we render free-body diagrams, kinematic visuals, and structured CoT traces capturing the full solution process.

We also leverage openly licensed resources such as OpenStax (MIT OpenCourseWare, 2022) and MIT OCW (Moebs et al., 2016) to generate more diverse and complex physics problems, ultimately achieving scalable and legally clear dataset generation while ensuring diverse, high-quality examples.

Chemistry. Organic reaction prediction is a classic multimodal reasoning task, typically framed as symbolic input and structural output. We include a chemistry subset of 4,700 two-to-one reactions from the **USPTO-50K** dataset (Ramsundar et al., 2019), filtered for distinct reactants and single products. Each reaction trace includes



Chemistry

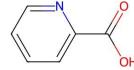
Question: When 4-chloro-3,5-dipyridin-3-yloxybenzene-1,2-diamine undergoes reaction with pyridine-2-carboxylic acid, what product is formed?

<think>**Thought 1:** To understand the reaction, let's first visualize the structure of the diamine reactant, 4-chloro-3,5-dipyridin-3-yloxybenzene-1,2-diamine. This will help identify the reactive amino groups. </think>



Reasoning Image 1

<think>**Thought 2:** The first sketch shows the diamine structure with two adjacent amino groups (-NH2) on the benzene ring, along with chloro and pyridinyloxy substituents. Now, let's visualize the second reactant, pyridine-2-carboxylic acid. </think>



Reasoning Image 2

three visual artifacts: individual molecular depictions of each reactant, a combined schematic of both reactants side-by-side, and the resulting product structure. Molecules are rendered with RDKit, and names are retrieved from PubChem when available. Text prompts use randomized templates (e.g., “What is formed by combining acetic acid and ethanol?”), and PubChem names are included when available. This visual progression helps models learn compositional chemical structure without SMILES or reaction templates.

Algorithmic problem solving. Humans naturally create visual diagrams when solving complex problems, transforming abstract concepts into spatial representations for deeper reasoning. We formalize this by interpreting coding problems through compact visual scaffolds: one or two diagrams depicting graph structure, edge weights, etc. To build traces, we run an iterative “visual sketchpad” loop: GPT-4.1 receives a prompt and returns THOUGHT statements plus VIS_SPEC blocks when sketches are needed; we render specs with networkx/matplotlib, feed images back to the model, and repeat until complete, then clean transcripts with post-processing.

Problem samples come from competitive programming, prioritizing real-world abstractions like logistics, network routing, and flow optimization. The orchestrator produces simple visual structures emphasizing clarity over style. Each trace contains the problem prompt, 1–3 reference diagrams, and polished explanations, supporting grounded reasoning in discrete structures while mirroring how algorithms are taught. The final corpus comprises 1,200 diverse algorithm-based problems spanning competitive programming.

Graph problems. Graph algorithms are useful for large language model applications because they efficiently organize and traverse structured relationships, for example in search and retrieval applications. Methods like shortest-path and subgraph matching enable multi-step reasoning by connecting relevant concepts across knowledge graphs. Recent work by Fu et al. (2024a) shows that although LLMs can solve graph problems such as connectivity and maximum flow to some extent when a textual description of the graph is given, *multimodal* LLMs suffer when solving graph problems. This finding suggests potential for improving multimodal models’ graph-understanding abilities by guiding their reasoning over images.

We create 10,000 graph problems with full reasoning traces spanning over four tasks: graph connectivity, shortest path, minimum spanning tree, and topological sort. Each task has about 2,500 samples, with one problem image and at most 19 reasoning images per sample. Each reasoning image is coupled with an explanation for the underlying algorithms, for example, Dijkstra for the shortest path, BFS for connectivity, etc.

4.2 2D Visual Reasoning

Visual search. Previous research has shown that drawing bounding boxes and zooming can improve accuracy on visual search tasks (Wu and Xie, 2024; Shao et al., 2024a). We follow such tasks by creating two types of traces, one for drawing bounding boxes and one for zooming. We use data from Shao et al. (2024a) to generate our traces covering four categories of visual search tasks: chart, text/doc, relation study, and general VQA.

Visual jigsaw. Visual jigsaw refers to filling in missing pieces of an image, as in a jigsaw puzzle. Each puzzle is constructed from an ImageNet (Deng et al., 2009) image, with 1 to 4 missing pieces of varying shapes, including rectangles and irregular regions. Each puzzle includes four multiple-choice options, where each option presents a set of candidate missing pieces. Only one set correctly matches the pieces removed from the original ImageNet image. We generate two types of visual CoT traces for solving each puzzle. In the first type, we iteratively fill in the missing patches using the pieces from each multiple-choice option and identify the one that produces a coherent



Visual Search

Question: What type of clothing is the boy in the foreground wearing while playing on the field?



Problem Image 1

<think>**Thought 0:** To determine the boy's clothing, I need to closely examine his attire in the provided image. **Thought 1:** I will zoom in on the boy to get a clearer view of his shirt, shorts, and legwear. </think>



Reasoning Image 1

<think>**Thought 2:** The zoomed-in image clearly shows he is wearing a white short-sleeved shirt with black trim, black shorts, and white socks with black stripes, typical of a soccer uniform. **Thought 3:** Based on these details, the boy is dressed in athletic wear suitable for playing soccer.</think>



Visual Jigsaw

Question: A jigsaw puzzle shows a picture of pandas but has several missing pieces, indicated by gray areas. Identify which set of pieces from the options provided will correctly complete the puzzle.



Problem Image 1

<think>**Thought 0:** Analyze the original puzzle image to understand the missing sections and the surrounding details. **Thought 1:** To determine the correct missing pieces, visualize the complete image or find the original source image. </think>



Reasoning Image 1

<think>**Thought 2:** Compare the full image with the options provided to see which set of pieces matches the content and position of the gray areas in the original puzzle. **Thought 3:** Option D contains the correct parts of the image needed to fill the missing sections based on the full image.</think>

image. In the second type, we imagine what the original image would look like and then select the option whose pieces best match the imagined reconstruction.

4.3 3D Visual Reasoning



Embodied CoT

Question:
Please describe the sequence of robot actions from the initial state (the first image) to the goal state (the second image).



Problem Image 1 & 2

<think>**Thought 1:** turn left, take a step, turn right, walk to the sink.</think>



Reasoning Image 1

<think>**Thought 2:** pick up the candle that is on the counter by the sink.</think>



Reasoning Image 2

<think>**Thought 3:** turn around, walk to the shower, turn left, walk to wall, turn left to face the toilet. </think>



Reasoning Image 3

<think>**Thought 4:** put the candle on the back of the toilet. </think>



Reasoning Image 4

Embodied planning. For embodied planning tasks, agents must *ground high-level decisions* in the evolving visual context of the environment. We reformulate the **ALFRED** (Shridhar et al., 2020) benchmark, an interactive 3D simulation environment where agents perform complex tasks based on human instructions, into an image goal-conditioned planning task.

In this new task, the model receives two images — the initial and goal states — and is tasked with generating a textual description of the high-level planning steps required to transition from the initial to the goal state. To

emphasize the role of visual reasoning, we require the generated descriptions to be detailed and step-by-step (e.g., “*turn and go to the TV; pick up the bowl that is on the TV stand in front of the TV; with the bowl in hand...*”) rather than brief summaries (e.g., “*move bowl to coffee table*”), which can often be produced through shortcut reasoning without capturing intermediate visual steps.

We compile the entire training set, as well as the seen and unseen validation sets from ALFRED, resulting in a total of 7,080 examples spanning diverse visual reasoning trajectories. When multiple textual reasoning annotations exist for a single visual trajectory, we include all of them, resulting in 22,666 textual reasoning traces.

Robot Planning

Question:
Please describe the sequence of robot actions from the initial state (the first image) to the goal state (the second image).

Problem Image 1 & 2

Reasoning Image 1

Reasoning Image 2

Reasoning Image 3

<think> **Thought 1:** move the robot's arm towards the bread.
</think>

<think> **Thought 2:** use the robot's arm to grab the bread.
</think>

<think> **Thought 3:** move the bread towards the table.
</think>

Robot planning. While low-level manipulation may rely on reactive control, continuous planning for complex tasks often requires *high-level visual guidance*, making visual CoT essential for bridging perception and long-horizon decision-making in robot planning. Similarly, we reformulate **RoboMIND** (Wu et al., 2024), a multi-embodiment dataset of real-world robot manipulation, into an image goal-conditioned planning task. In this setting, a model is provided with the initial and goal states images, along with a textual description of the robot setup (e.g., AgileX (AgileX Robotics, 2023), Franka (Franka Emika GmbH, 2018), or UR5e (Universal Robots A/S, 2018)), and is tasked with generating a detailed textual plan outlining the high-level steps required to transition from the initial to the goal state.

Unlike embodied planning tasks that often involve partial observability and require agents to infer unobserved states, this robot planning task is fully observable. Therefore, the challenge lies not in imagining the visual trajectory but in articulating precise movements for each arm or gripper to accomplish the task (e.g., “[left] move towards the oven door and [right] grab the corn.”).

To control degrees of freedom, we exclude the humanoid robot examples from the original RoboMIND dataset, focusing solely on tasks involving robotic arms. This results in a curated subset of 6,945 robot planning tasks, each annotated with human-generated high-level actions that serve as visual reasoning trajectories.

3D Multi Hop Object Counting

Question: Remove zero CAPSULE objects. Eliminate all PYRAMID objects. Remove zero PURPLE objects. Discard all CUBE objects. Count how many blue-colored objects remaining.

Problem Image 1

Reasoning Image 1

<think> **Thought 2:** Proceeding to Step 2, the instruction is to “Remove all PYRAMID objects.” Upon reviewing the current scene, there is one orange pyramid present. This object must be removed from the arrangement. The updated scene after this removal is depicted in the following image.</think>

<think> **Thought 3:** For Step 3, the directive states to “Remove 0 PURPLE object.” Similar to Step 1, this operation specifies that no purple objects are to be removed. The objects remaining in the scene after Step 2 are preserved... </think>

3D multi-hop objects counting. A core aspect of human visual-spatial reasoning is understanding transformations and imagining scenes from different viewpoints. For this task, our setup follows a structure similar to that of Johnson et al. (2017), using 10 predefined shape types (e.g., sphere, cylinder, donut) in various colors. At each step, we randomly apply one of three operations: remove all instances of an attribute (e.g., all red objects), remove a subset (e.g., 5 red objects), or add new objects (e.g., 2 blue prisms, 1 red sphere). We then create questions that ask about the quantity of specific attributes or what objects are left in the field. To increase difficulty, the initial scenes are rendered from varying viewpoints (front, back, left, right), where some objects may be partially occluded by those in front. The first visual reasoning step involves generating a top-down 45° view to reconstruct the full scene, allowing the model to see potentially blocked objects. The subsequent visual sketches correspond to each transformation step in the instruction. We also improve upon the data from Johnson et al. (2017) by adding in different materials, backgrounds, and floor designs.

4.4 Visual Logic and Strategic Games

Mazes. Mazes serve as a canonical testbed for visual CoT reasoning, bridging low-level perception with high-level symbolic search. Unlike purely pixel-based 2D visual tasks such as visual search and visual jigsaw, mazes possess explicit graph structure yet remain visually intuitive, letting us disentangle vision errors from planning errors.

We adopt the [maze-dataset](#) library to procedurally generate thousands of grid mazes with diverse topologies (lattice type, branch factor, loop density).¹ Each instance is exported in two complementary formats: a) `m.as_pixels()`, an RGB raster that encodes walls, free cells, start ■, and goal ■, suitable for visual perception; b) `MazePlot`, a vector overlay that can superimpose solution paths, candidate trajectories, heat-maps, or landmark nodes for human-readable walk-throughs. To increase maze diversity, we also use OpenAI Gym’s FrozenLake-v1 environment (Brockman et al., 2016).

We evaluate a broad spectrum of spatial reasoning skills across multiple question types: *topological analysis, pathfinding, navigation planning, and coverage problems*. Maze complexity varies across multiple configurable parameters to ensure tasks require genuine planning rather than memorization. More details in [Appendix A.3](#).

Chess. Strategic planning in chess involves simulating multiple futures and selecting moves that maximize long-term advantage. To support counterfactual reasoning, we construct a dataset of mid-game positions from rated Lichess games², each with structured visual traces. Given a position, Stockfish identifies the optimal move, and three alternates are sampled randomly from legal moves. Each candidate is visualized independently for comparative evaluation. By rendering possibilities in isolation, move consequences—tempo gain, structural weakening, tactical motifs—become legible, enabling better strategic reasoning. Traces are formatted as multiple-choice tasks with visual sketches, encouraging tactical awareness and pattern recognition. Postprocessing with Gemini 2.5 Flash refines traces into coherent multimodal reasoning sequences for short and long-horizon planning, yielding 11,500 chess traces across diverse positions and hypothetical move states.

Visual logic puzzles. Humans approach logic puzzles such as Tetris, Raven’s Progressive Matrices (RPM, Zhang et al., 2019), and the Abstraction & Reasoning Corpus (ARC-AGI, Chollet, 2019; Chollet et al., 2024) primarily through visuospatial reasoning: we see how pieces combine, transform, or complete a pattern before committing

¹maze-dataset supports recursive-backtracker, randomized Prim, Wilson, and Kruskal generators; see (Ivanitskiy et al., 2023).

²<https://lichess.org/>



Chess

Question: What is the best move for Black to play?
 A: Ne7
 B: Kf8
 C: Qe7
 D: Qe6



Problem Image 1

<think>**Thought 1:** The initial position shows Black's king on e8, potentially vulnerable, and the knight on f5 is pinned to the queen on d7 by White's bishop on g5. Black needs to find a move that improves the position or defends against potential threats. Let's evaluate the candidate move Ne7. This move breaks the pin on the knight.</think>



Reasoning Image 1

<think>**Thought 2:** After Ne7, the knight is no longer pinned, but it is still attacked by the queen on e4 and also blocks the queen's defense of the d6 pawn. This move doesn't seem to improve Black's position significantly...</think>

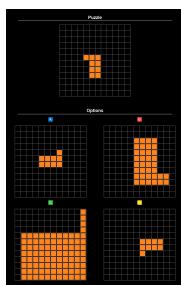
to an answer. These logic games rely heavily on visuospatial working memory, which is correlated with general intelligence level (Lau-Zhu et al., 2017; de Winter et al., 2023).

To enhance models with such cognitive ability, we include the following tasks. For *Tetris*, we collect three types of tasks: a) shape assembly: given a silhouette and candidate tetromino sets, select the one that perfectly tiles the shape; b) grid completion: fill a partially occupied grid using a specified set of tetrominoes; c) spatial transformation: apply a sequence of geometric operations (translate, rotate, mirror, scale) to an irregular shape in the grids. The visual CoT involves visualizing each transformation step. For *RPM* (IQ matrix), we include three types from Zhang et al. (2019) that involve compositional reasoning. The reasoning trace identifies visual patterns for each compositional component across rows or columns. For *ARC-AGI*, while prior models often rely on textual reasoning, humans typically solve these tasks through visual pattern recognition. To better align with human strategies, we construct two types of visual CoT. The first begins with matrix representations of the training examples and test input; the reasoning trace first visualizes the training examples, the test input, and finally the predicted output. The second type directly uses visual representations in the task instruction, thus the model only has to generate a visual sketch of the predicted output as part of its reasoning process. For all data, we use VLM to generate accompanying textual descriptions to enrich interleaved text-image rationales.



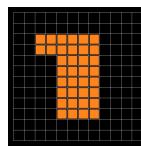
Tetris

Question:
 Apply the following sequence of transformations to the blue shape: scale by 2 X, then rotate 90 °clockwise, then translate 1 up, then rotate 90 °clockwise. Choose the option that shows the resulting shape.



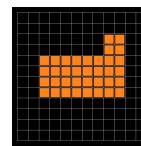
Problem Image 1

<think>**Thought 1:**
 Apply the first transformation: scale the shape by 2 times. This expands the shape.</think>



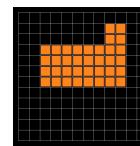
Reasoning Image 1

<think>**Thought 2:**
 Apply the second transformation: rotate the scaled shape 90 degrees clockwise. </think>



Reasoning Image 2

<think>**Thought 3:**
 Apply the third transformation: translate the rotated shape 1 unit upwards. </think>



Reasoning Image 3

Checkers. Our checkers traces, initiated from randomized mid-game boards and continued through depth-4 minimax, are designed to capture key tactical motifs such as forced captures, multi-jump chains, and king-row advancement, providing dense supervision for spatially grounded decision making. To ensure data quality, we apply rule-based validators that discard invalid traces. Final traces are passed through a language model to synthesize coherent visual reasoning aligned with *turn-based planning*, ending with 2,800 full traces.

Ciphers. Classical encryption schemes convert symbolic rules into spatial transformations, enabling visual reasoning over grids and coordinates. We generate examples of *Scytale*, *Polybius*, *Rail Fence* and *Playfair* ciphers, each

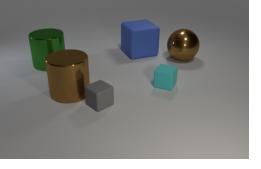
with multimodal traces showing encoding steps. Visual sketches highlight relevant regions—column positions, grid lookups, or digraph geometry—while text describes applied rules (e.g., “Letter ‘P’ goes on the second rail,” “‘A’ is at coordinates (1, 1).”) Early steps use visuals, later steps use language, mirroring real cryptanalysis. The 6,600 traces align symbolic manipulation with visual structure.

Connect 4. We also include 2,100 Connect-Four traces that teach short-horizon grid planning. Each starts from a mid-game position obtained after 4–20 random moves; a depth-4 α - β minimax (center-control, open-three heuristic) then plays the forced win. Every drop is rendered on a 7×6 board with a glowing overlay, and an LLM condenses the raw log into a clear, step-by-step “How does Red win?” narrative pairing text and visuals.

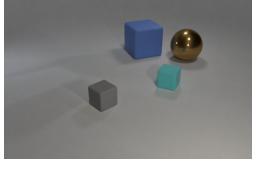


Example Interleaved Visual Reasoning Chains Generated by 🍞 Bagel-Zebra-CoT

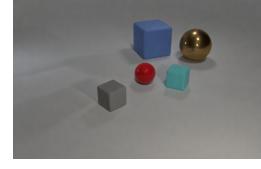
Question: Subtract all cylinders. Add 1 red sphere. How many objects are left?



Problem Image 1

<think>
Remove all the cylinders.


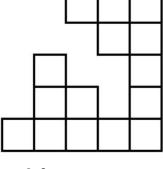
Reasoning Image 1

<think>
Add 1 red sphere.


Reasoning Image 2

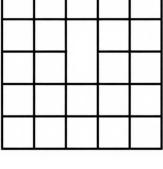
<answer>
Final Answer: 5

Question: A big square is made from 25 small squares put together. A few of the small squares have been lost. How many have been lost?



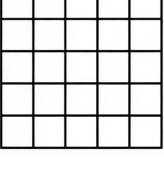
Problem Image 1

<think> The problem asks for the number of missing small squares in a 5x5 grid. We can see the complete 5x5 grid structure. To find the missing squares, we need to see the full pattern. The first image shows the complete grid.</think>



Reasoning Image 1

<think> Comparing the complete grid in the previous image with the given puzzle image, we can count the number of squares that are present in the complete grid but missing in the puzzle. By counting the missing squares in the complete grid, we find the total number of lost small squares.</think>



Reasoning Image 2

<answer>
Final Answer: 10

Figure 4 Example interleaved reasoning chains generated by Bagel-Zebra-CoT, a Bagel-7B model finetuned on ZEBRA-COT. These traces demonstrate ZEBRA-COT’s for instilling intrinsic visual reasoning capability in complex multimodal models.

5 Training Models on ZEBRA-COT

Anole-Zebra-CoT. We fine-tune Anole (Chern et al., 2024) on our dataset, which builds on Chameleon (Team,

Model	MathVision	MathVista	VisuLogic	EMMA	MMVP	Blink	Vstar
Anole with CoT prompting	13.80	22.80	8.50	12.80	10.00	26.46	23.60
Anole-Zebra-CoT (Ours)	16.45	25.30	21.80	15.02	15.33	31.25	27.20

Table 2 Overall performance (%) across eight datasets for the base Anole model with chain-of-thought prompting v.s. the same Anole model further trained upon our ZEBRA-COT data. A full breakdown of each evaluation set is presented in [Appendix C](#).

2024), using the codebase from Chern et al. (2025). We finetune the model fully end-to-end on a node with $8 \times$ H200 GPUs for 12 hours, with a learning rate of 1×10^{-5} , cosine decay, a batch size of 8, and a max token length of 12288. We train the model for 10k steps. To evaluate our trained model, we set the max generation length to 16384. After fine-tuning Anole on our Zebra-CoT corpus, the accuracy goes up from 4.2% (6 / 142) to 16.9% (24 / 142), delivering a 4 times relative performance improvement and a 12% gain in accuracy.

Furthermore, we evaluate on seven challenging benchmarks that require visual reasoning, including MathVision (Wang et al., 2024a), MathVista (Lu et al., 2024), VisuLogic (Xu et al., 2025a), EMMA (Hao et al., 2025b), MMVP (Tong et al., 2024), BLINK (Fu et al., 2024b), and Vstar (Wang et al., 2023). All the evaluations are done using VLMEvalKit (Duan et al., 2024). To ensure a fair comparison, we use chain-of-thought prompting (Wei et al., 2022) when evaluating the base Anole model. As shown in [Table 2](#), training with ZEBRA-COT significantly improves the Anole model across all benchmarks. Most notably, it could improve the Anole model’s visual logical reasoning capabilities by 13.3 points.

Bagel-Zebra-CoT. To further test whether ZEBRA-COT can enhance a stronger backbone, we fine-tune the BAGEL-7B model (Deng et al., 2025) end-to-end on a node with $8 \times$ H200 GPUs for 1,000 steps using packed sequences with 60,000 tokens with a learning rate of 2×10^{-5} and cosine decay. We cap all images at a resolution of 512 of the min side, resulting in approximately 1,024+ visual tokens per image. Because the original Bagel implementation cannot natively generate interleaved text-image outputs, we revise the training loop to include a loss term at the `<|vision_start|>` token, enabling seamless visual token generation. We additionally wrap text reasoning tokens between `<think>` and `</think>`, and the final answer within `<answer>` and `</answer>`. At inference time, when encountering `<im_end>`, we sample one additional token to check whether the next token is `<|vision_start|>`; if so, the model itself seamlessly switches to image generation mode to generate visual aids. The entire interleaved generation process only stops if the model generates the `<answer>` token.

We observe that our trained model can inherently generate visual CoT when solving problems, even on tasks outside its training distribution. This suggests its potential as a strong initialization for future reinforcement learning fine-tuning. In [Figure 4](#), we include representative reasoning traces produced by the model.

6 Conclusion & Future Directions

In this paper, we introduced ZEBRA-COT, a large-scale dataset of 182K interleaved text-image reasoning traces spanning 4 major categories across 18 domains with over 50 distinct tasks. Fine-tuning experiments demonstrate substantial improvements: Anole-7B achieves an average 4.9 % gain across seven challenging benchmarks, with up to 13.1% on visual logic tasks, while Bagel-7B learns to inherently generate visual aids during problem solving, a capability absent in the base model.

This work opens several exciting avenues for future research. Most immediately, models trained on ZEBRA-COT, particularly our Bagel variant that natively generates visual thoughts, provide strong initializations for reinforcement learning. Just as text-based reasoning models have benefited from RL fine-tuning to improve logical consistency and correctness, we envision similar gains for visual reasoning through RL with verifiable rewards (Shao et al., 2024b; Guo et al., 2025) or fine-grained rewards (Zeng et al., 2024; Fu et al., 2025).

We believe ZEBRA-COT represents a crucial step toward AI systems that think visually as naturally as humans sketch diagrams, generate graphs, and use spatial reasoning to solve complex problems. By open-sourcing our dataset and model weights, we hope to accelerate progress toward this goal.

Acknowledgments

MG and AL were supported by a Research Award from the Columbia Center of AI Technology in collaboration with Amazon and by Google. KY, ZC, FH, and TG were supported by DARPA Transfer from Imprecise and Abstract Models to Autonomous Technologies (TIAMAT) 80321, National Science Foundation NSF-IIS-2147276 FAI, and DOD-AFOSR-Air Force Office of Scientific Research under award number FA9550-23-1-0048. RJ was supported by the National Science Foundation under Grant No. IIS-2403436. VS was supported in part by an NSF CAREER Award CCF-2239265 and an Amazon Research Award. WN was supported in part by the National Science Foundation under Grant No. CMMI-2427856. OL, DF, and BZ would like to thank Center for Advanced Research Computing (CARC) and NLP Group at USC for providing compute resources. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

- AgileX Robotics. Cobot magic. <https://global.agilex.ai/products/cobot-magic>, 2023. 8
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. 2022. URL <https://arxiv.org/abs/2204.14198>. 3
- Avinash Anand, Janak Kapuriya, Apoorv Singh, Jay Saraf, Naman Lal, Astha Verma, Rushali Gupta, and Rajiv Shah. Mm-phyqa: Multimodal physics question-answering with multi-image cot prompting. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 53–64. Springer, 2024. 3
- Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Guha, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, et al. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *Advances in Neural Information Processing Systems*, 37:36805–36828, 2024. 2, 4
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 3
- Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016. 9
- Xi Chen and Xiao Wang. Pali: Scaling language-image learning in 100+ languages. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3
- Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024. 1, 2, 11
- Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images. *arXiv preprint arXiv:2505.22525*, 2025. 1, 2, 12
- François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. 9
- François Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. Arc prize 2024: Technical report. *arXiv preprint arXiv:2412.04604*, 2024. 9
- Joost CF de Winter, Dimitra Dodou, and Yke Bauke Eisma. Responses to raven matrices: Governed by visual complexity and centrality. *Perception*, 52(9):645–661, 2023. 10
- Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 1, 2, 12
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6
- 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. 12
- Franka Emika GmbH. *Franka Emika Panda Robot Arm*, 2018. <https://www.franka.de>. 8

Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. IsoBench: Benchmarking multimodal foundation models on isomorphic representations. In *First Conference on Language Modeling (COLM)*, 2024a. 6

Deqing Fu, Tong Xiao, Rui Wang, Wang Zhu, Pengchuan Zhang, Guan Pang, Robin Jia, and Lawrence Chen. TLDR: Token-level detective reward model for large vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=Zy2XgaGpDw>. 12

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. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024b. 3, 12

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 12

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhui Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*, 2024. 3

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025a. 3

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025b. 12

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL <https://openreview.net/forum?id=7Bywt2mQsCe>. 5, 18, 25

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*, 2024. 1, 2, 5

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 3

Michael Igorevich Ivanitskiy, Rusheb Shah, Alex F. Spies, Tilman Räuker, Dan Valentine, Can Rager, Lucia Quirke, Chris Mathwin, Guillaume Corlouer, Cecilia Diniz Behn, and Samy Wu Fung. A configurable library for generating and manipulating maze datasets, 2023. URL <https://arxiv.org/abs/2309.10498>. 9

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*, pages 2901–2910, 2017. 3, 9

Alex Lau-Zhu, Emily A Holmes, Sally Butterfield, and Joni Holmes. Selective association between tetris game play and visuospatial working memory: A preliminary investigation. *Applied cognitive psychology*, 31(4):438–445, 2017. 10

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023. 3

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025. 1, 3, 5

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024a. 3

Qingyun Li, Zhe Chen, Weiyun Wang, Wenhui Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. *arXiv preprint arXiv:2406.08418*, 2024b. 2

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 3

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024. 12

Sachit Menon, Richard Zemel, and Carl Vondrick. Whiteboard-of-thought: Thinking step-by-step across modalities. *arXiv preprint arXiv:2406.14562*, 2024. 2

MIT OpenCourseWare. [Course Title]. <https://ocw.mit.edu/>, 2022. MIT OpenCourseWare: Massachusetts Institute of Technology. License: Creative Commons BY-NC-SA. 5

Chanchik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024. 2

William Moebs, Samuel J. Ling, and Jeff Sanny. *University Physics Volume 1*. OpenStax, Houston, Texas, 2016. URL <https://openstax.org/books/university-physics-volume-1/pages/1-introduction>. Licensed under CC BY 4.0. 5

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhui Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36:25081–25094, 2023. 3

OpenAI. Openai o3 and o4-mini system card. Technical report, OpenAI, April 2025a. URL <https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf>. 1

OpenAI. Thinking with images. <https://openai.com/index/thinking-with-images/>, April 2025b. Accessed: 2025-07-21. 1, 2

Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu. *Deep Learning for the Life Sciences*. O'Reilly Media, 2019. <https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837>. 5

Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chimmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*, 2023. 2

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024a. 1, 2, 3, 5, 6

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024b. URL <https://arxiv.org/abs/2402.03300>. 12

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL <https://arxiv.org/abs/1912.01734>. 7

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. *arXiv preprint arXiv:2502.13383*, 2025. 3

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 1, 3

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898, 2023. 1, 2

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 1, 11

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovićova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 3

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. 12

Universal Robots A/S. *UR5e Collaborative Robot Arm*, 2018. <https://www.universal-robots.com/products/ur5e/>. 8

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL <https://openreview.net/forum?id=QWTCcxMpPA>. 12

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b. 3

- Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Videocot: A video chain-of-thought dataset with active annotation tool. *arXiv preprint arXiv:2407.05355*, 2024c. 3
- Yuxuan Wang, Zilong Zheng, Xueliang Zhao, Jinpeng Li, Yueqian Wang, and Dongyan Zhao. VSTAR: A video-grounded dialogue dataset for situated semantic understanding with scene and topic transitions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5036–5048, Toronto, Canada, July 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.acl-long.276>. 12
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 12
- Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhiqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, Zhen Zhao, Guangyu Li, Zhao Jin, Lecheng Wang, Jilei Mao, Xinhua Wang, Shichao Fan, Ning Liu, Pei Ren, Qiang Zhang, Yaoxu Lyu, Mengzhen Liu, Jingyang He, Yulin Luo, Zeyu Gao, Chenxuan Li, Chenyang Gu, Yankai Fu, Di Wu, Xingyu Wang, Sixiang Chen, Zhenyu Wang, Pengju An, Siyuan Qian, Shanghang Zhang, and Jian Tang. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024. 8
- Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024. 1, 2, 5, 6
- Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. 3
- Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhui Wang, Jifeng Dai, and Jinguo Zhu. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models, 2025a. URL <https://arxiv.org/abs/2504.15279>. 12
- Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*, 2025b. 1, 3
- Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025. 3
- Zhengyuan Yang*, Linjie Li*, Jianfeng Wang*, Kevin Lin*, Ehsan Azarnasab*, Faisal Ahmed*, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. 2023. 2
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 3
- Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 58348–58365. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/zeng24c.html>. 12
- Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327, 2019. 9, 10
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 2
- Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. *arXiv preprint arXiv:2503.22020*, 2025. 3
- Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. 1, 2
- Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36:8958–8974, 2023. 2, 3

A Dataset Details

A.1 Data Statistics.

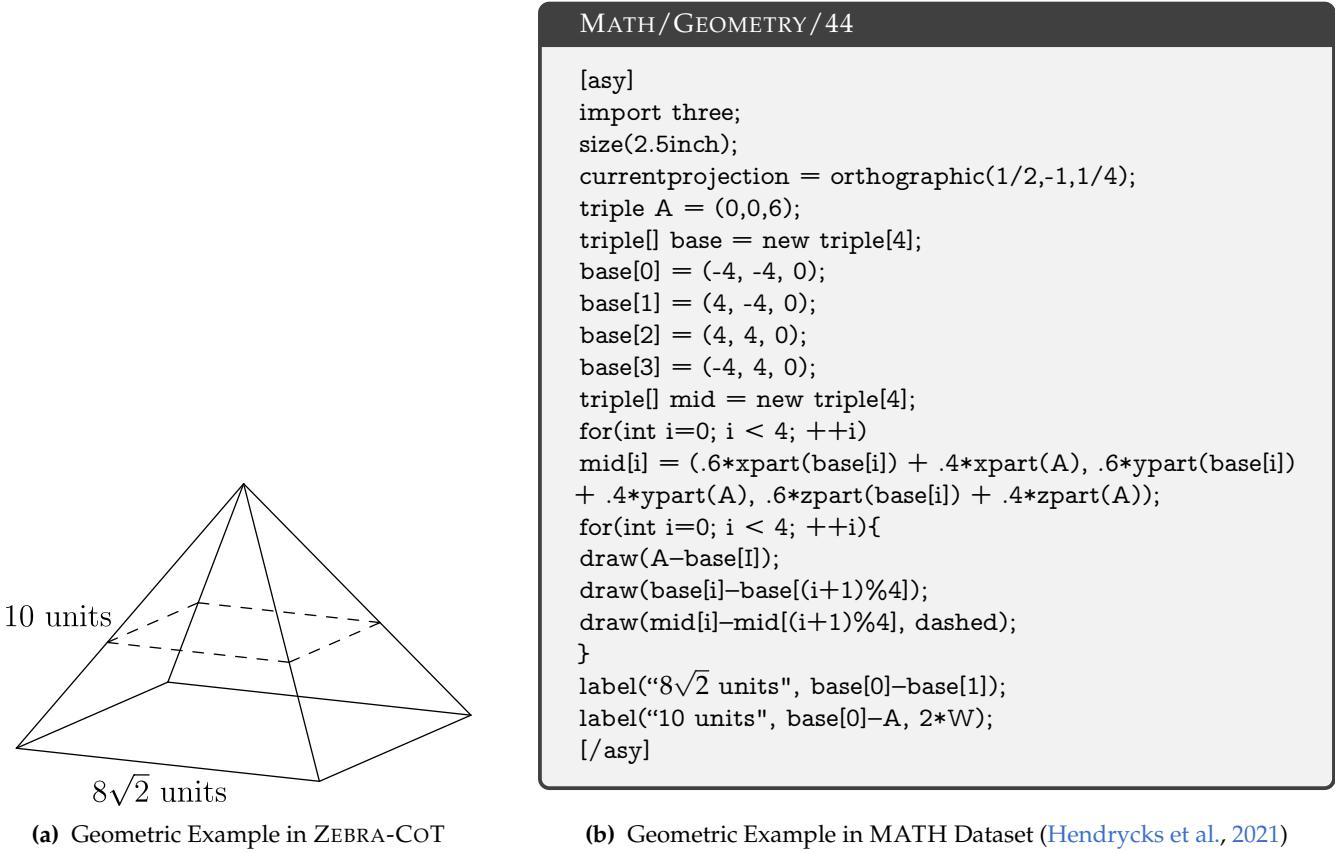
Here we show detailed statistics about ZEBRA-COT’s categories.

Table 3 Statistics of ZEBRA-COT.

General Category	Sub Category	Count	Percentage (%)
2D Visual Reasoning	Visual Jigsaw	21,899	12.0
	Visual Search	30,000	16.4
	Subtotal	51,899	28.5
3D Visual Reasoning	Embodied Cot	22,666	12.4
	Multi-Hop Objects Counting	10,000	5.5
	Robot Planning	6,944	3.8
	Subtotal	39,610	21.7
Scientific Reasoning	Chemistry	4,666	2.6
	Competitive Programming	1,207	0.7
	Geometry	1,058	0.6
	Graph Algorithms	10,000	5.5
	Physics	7,090	3.9
	Subtotal	24,021	13.2
Visual Logic Strategic Games	Arc-Agi	2,000	1.1
	Checkers	2,753	1.5
	Chess	20,483	11.2
	Ciphers	6,589	3.6
	Connect Four	2,029	1.1
	Maze	20,000	11.0
	RPM	3,000	1.6
	Tetris	10,000	5.5
	Subtotal	66,854	36.7
Total		182,384	100.0

A.2 Math Geometry Details.

Here, we provide example code for geometry sketch generation.



(a) Geometric Example in ZEBRA-COT

(b) Geometric Example in MATH Dataset ([Hendrycks et al., 2021](#))

Figure 5 Comparison of the same geometric figure in our ZEBRA-COT dataset and the MATH dataset. Ours focus on multimodal reasoning and explicitly plot the geometry problem than using the text-only plotting codes.

A.3 Maze Details

We evaluate a broad spectrum of spatial reasoning skills across multiple question types: *I. topological analysis* (e.g., counting isolated regions, identifying connected components under 4- or 8-connectivity, finding the largest connected area), *II. pathfinding* (e.g., determining reachable endpoints, computing shortest paths, enumerating all optimal routes), *III. navigation planning* (e.g., selecting correct paths from alternatives, calculating minimal moves to reach targets), and *IV. coverage problems* (e.g., visiting all marked locations, identifying the farthest reachable position). This diverse task suite goes beyond simple start-to-goal navigation, encompassing the full range of spatial reasoning strategies that humans use to interpret complex environments. We also introduce varying complexity of the matrix, including different maze side lengths ranging from $(5, 15)$, different branching factors b , loop probability ℓ , and number of distractor endpoints k . Larger n exponentially increases the search space, while higher b and ℓ degrade heuristic admissibility. Both of those require genuine planning rather than rote memorization.

B Prompt Templates

B.1 Prompt for Enhancing Raw Reasoning Traces for Online and Agentic Data

Prompt Template 1

You are an expert in creating clean and logically coherent multimodal chain of thought traces. Your
→ task is to analyze
and comprehend a raw reasoning trace with interleaved text and images, then transform it into a
→ clean, step-by-step multimodal
reasoning trace that correctly solves the original problem.

- ===== INPUT =====
1. Problem & Noisy Trace: A raw interleaved text and image reasoning trace. Images in this trace are
 - represented by placeholders:
 - `'[problem_image_X]'` for original problem images (e.g., `'[problem_image_1]'`,
→ `'[problem_image_2]'`)
 - `'[reasoning_image_X]'` for images generated during reasoning (e.g., `'[reasoning_image_1]'`,
→ `'[reasoning_image_2]'`)
 2. Image Data: The actual image data corresponding to the placeholders, provided separately.

===== Your Task =====
Generate a clean, logical multimodal reasoning trace as **plain text** that represents the *ideal*
→ reasoning process to solve the problem.

===== OUTPUT FORMAT =====
You MUST generate the formatted reasoning trace with the following structure:

QUESTION:

<The original problem statement with text and image placeholders:
→ <image_start>[problem_image_1]<image_end>,
→ <image_start>[problem_image_2]<image_end>, etc. Stay as close to the original problem
→ statement as possible but remove noise to ensure clarity>

REASONING TRACE:

THOUGHT 0: <Clear description of initial reasoning step that identifies key elements of the
→ problem>
THOUGHT 1: <Next reasoning step, often explaining why an image will be created>
<image_start>[reasoning_image_1]<image_end>
THOUGHT 2: <Further reasoning step based on the image, explaining insights gained>
<image_start>[reasoning_image_2]<image_end>
// Additional thoughts and images as needed
<image_start>[reasoning_image_X]<image_end>
THOUGHT N: <Final reasoning step before the answer, summarizing key insights>

FINAL ANSWER:

<The final calculated answer based on the reasoning>

===== Guidelines =====

1. Enhancing Original Trace Rather than Generating New Trace:

- Instead of generating a new trace, your task is to enhance the original trace (which is a correct
→ trace but rather concise and lacks coherent multimodal reasoning) by adding more details and
→ explanations, see the following sections of guidelines for more details.
- You MUST use all the images provided in the original trace.

- You should use the original trace as a reference rather than copying it verbatim.
2. Multimodal Reasoning Flow:
- Develop a coherent, step-by-step chain of thought that seamlessly integrates textual and visual reasoning.
 - Clearly explain the necessity of generating a sketch / visual thought / image before introducing its placeholder.
 - After each image placeholder, describe the insights gained from the sketch / visual thought / image, and how it contributes to advancing the solution.
 - Ensure each step logically builds on the previous ones, especially between text reasoning and visual reasoning steps.
3. Image Placeholders and References:
- Use placeholder tags ONLY when you want to actually insert/show/generate an image in your trace. When doing so, write the corresponding placeholder tag exactly as shown, including the <image_start> and <image_end> tags.
 - Each unique image in the original problem and the reasoning trace should be represented by a unique placeholder tag, and each unique placeholder tag should only show up once in the trace.
 - When referring to images in your explanations, use natural language descriptions (e.g., "the diagram in the question", "the first sketch", "the visual thought X I created") instead of using placeholder tags. This is important because it helps us to parse into interleaved text and image sequences.
 - For images from the original problem, use: <image_start>[problem_image_X]<image_end>
 - For sketches or visuals generated during reasoning, use:
 - <image_start>[reasoning_image_X]<image_end>
4. Narrative Style:
- Remove irrelevant technical details such as debugging info, code snippets, and LaTeX package imports.
 - Eliminate verbose language that do not contribute to solving the problem.
 - Focus on the essential reasoning path that leads to the correct solution, using concise and clear language to describe the overall reasoning process.

B.2 Prompt for Enhancing Program Generated Template Data

Prompt Template 1

You are an expert in enhancing multimodal reasoning traces. Your task is to transform a template reasoning trace into a diverse multimodal reasoning trace that correctly solves the problem, while staying close to the original template and final answer.

===== INPUT =====

1. Problem & Template Trace: A template with interleaved text and image placeholders:

- `'[problem_image_X]'` for original problem images (e.g., `'[problem_image_1]'`)
- `'[reasoning_image_X]'` for images generated during reasoning (e.g., `'[reasoning_image_1]'`)

2. Image Data: The actual image data corresponding to the placeholders, provided separately.

===== Your Task =====

Generate a concise multimodal reasoning trace as **plain text**.

===== OUTPUT FORMAT =====

You MUST generate the formatted reasoning trace with the following structure:

QUESTION:

<Rewrite the problem statement in your own words while maintaining all key information. Do not
→ change key information. Include image placeholders:
→ <image_start>[problem_image_1]<image_end>,
→ <image_start>[problem_image_2]<image_end>, etc.>

REASONING TRACE:

THOUGHT 0: <Identify key elements of the problem>
THOUGHT 1: <Explain reasoning step, often why an image / sketch / visual thought is needed>
<image_start>[reasoning_image_1]<image_end>
THOUGHT 2: <Explain insights from the image>
<image_start>[reasoning_image_2]<image_end>
// Additional thoughts and images as needed
<image_start>[reasoning_image_X]<image_end>
THOUGHT N: <Summarize key insights before answer>

FINAL ANSWER:

<The original final answer in the template, do not change it>

===== Guidelines =====

1. Diversifying the Template:

- Rewrite the problem statement and reasoning steps in your own words while preserving all key information.
- Avoid deviating from the original template reasoning structure. Your job is to diversify the text of the original trace, not the logic.
- Vary the language and phrasing to avoid repetitive patterns.
- You MUST use all the images provided in the original trace.
- You MUST keep the original final answer.
- Maintain the original template's core reasoning structure and rationale while introducing textual refinements rather than substantial changes to the logical flow.

2. Multimodal Reasoning Flow:

- Develop a coherent, step-by-step chain of thought that seamlessly integrates textual and visual reasoning.
- Clearly explain the necessity of generating a sketch / visual thought / image before introducing its placeholder.
- After each image placeholder, describe the insights gained from the sketch / visual thought / image, and how it contributes to advancing the solution.
- Ensure each step logically builds on the previous ones, especially between text reasoning and visual reasoning steps.

3. Image Placeholders and References:

- Use placeholder tags ONLY when you want to actually insert/show/generate an image in your trace. When doing so, write the corresponding placeholder tag exactly as shown, including the <image_start> and <image_end> tags.
- Each unique image in the original problem and the reasoning trace should be represented by a unique placeholder tag, and each unique placeholder tag should only show up once in the trace.
- When referring to images in your explanations, use natural language descriptions (e.g., "the diagram in the question", "the first sketch", "the visual thought X I created") instead of using placeholder tags. This is important because it helps us to parse into interleaved text and image sequences.
- For images from the original problem, use: <image_start>[problem_image_X]<image_end>
- For sketches or visuals generated during reasoning, use:
→ <image_start>[reasoning_image_X]<image_end>

B.3 Prompt for Algorithmic Problems

Prompt Template 2

You are an expert in mathematical problem solving, algorithmic reasoning, visual explanation, and
→ creating multimodal reasoning traces.

1. STRICT VISUALIZATION POLICY (IMPORTANT):

You are only allowed to produce at most 3 [VIS_SPEC] visualizations, and they must all appear at
→ the very beginning of your reasoning (within the first 3–4 thoughts). You may only use the
→ following types for these visualizations:

- graph
- flow_network
- tree_from_dict
- tree_from_root
- grid

After these initial visualizations, you must do all further reasoning purely mentally/textually or with
→ pseudocode—NO MORE [VIS_SPEC] blocks are allowed after the first 3. Any attempt to
→ include more than 3 visualizations or use a disallowed type will be ignored.

The visual reasoning should only be used to understand the setup of the question - humans visualize
→ at the beginning to “set the board.” The actual problem solving is done purely textually.

General Rules:

- Interleave THOUGHT steps and [VIS_SPEC] image requests.
- Your final reasoned solution must match the logic of the given solution code.
- Prefix THOUGHT 0 with REASONING TRACE in the previous line.
- Prefix each reasoning step with “THOUGHT n.” (n starts at 0, less than 50 words each).
- Max 3 [VIS_SPEC] blocks, all within the first 3–4 thoughts.
 - Diagram #1: raw structural sketch (graph topology, blank grid, etc.).
 - Diagram #2–3: showcase pivotal elements if helpful.
- **Internal self-check (no output):** “Would a human scribble this as a quick setup sketch?” If the
→ answer is no, **do not** emit a VIS_SPEC.
- Strictly do not regenerate the same image - simply refer to the previous images in text if needed.
- Max of 10 thoughts.
- Every visualization request **must** use a minimal [VIS_SPEC] block with the correct type
→ specified. Do not use any other format.
- Do **not** include file names, imports, or drawing code. The orchestrator will handle image
→ generation.
- If you cannot meaningfully visualize or correctly visualize a thought using the provided tools and
→ inputs, then do not generate an image.
- Images are meant to be simple and visually cohesive - do not make grandiose images with titles and
→ axis - it's simply for a baseline understanding of the question.
- The first line of the trace should be QUESTION: followed by a detailed in depth recap of the
→ problem, specifying all the important aspects, without mentioning the solution.

2. Validity Rules:

- All [VIS_SPEC] parameters must be valid, fully-formed Python literals.
- For [VIS_SPEC] type "grid", the values must be a valid Python list of lists with exactly rows rows
→ and cols columns (or a flat list of length rows * cols), and each value should be a number or
→ string.
- For type graph, tree_from_dict, tree_from_root, and similar, node and edge labels may be strings
→ or integers, but all structures must be valid Python literals.

- Never output incomplete or empty lists/arrays/dicts in [VIS_SPEC] blocks. All lists must be fully
 - closed and contain at least one value, unless an empty structure is explicitly required by the problem.
 - Do not use variable names, symbolic labels, ellipses, or placeholders (e.g., a1, x, ..., an) anywhere in the [VIS_SPEC].
-

[VIS_SPEC] Reference Examples: Your blocks must follow the same format as these.

```
[VIS_SPEC]
type: graph
nodes: [A,B,C]
edges: [(A,B),(B,C)]
[/VIS_SPEC]
```

```
[VIS_SPEC]
type: flow_network
nodes: [A,B,C]
edges: [(A,B),(B,C)]
flows (optional): {(A,B): 2, (B,C): 1}
capacities (optional): {(A,B): 3, (B,C): 2}
[/VIS_SPEC]
```

...
...
...

3. Reflection step immediately after each VIS_SPEC

- Write a new THOUGHT that:
 - a. Describes what you see in the previous generated `reasoning_image_N.png`.
 - b. Explains how it informs your next reasoning move.

4. FINAL ANSWER

- After all reasoning, output “FINAL ANSWER.” and your concise solution (pseudocode is sufficient)

5. Formatting and Output Requirements

- Everything must be plain text with only the full QUESTION (just the problem itself, not the
 - name of the problem), FINAL ANSWER, REASONING TRACE marker, THOUGHT lines and
 - VIS_SPEC markers.

C Anole-Zebra Performance Breakdown

Split	Anole	Anole-Zebra-CoT (Ours)
Overall	12.80	15.03
Chemistry	12.84	15.48
Coding	9.75	16.31
Math	13.12	14.35
Physics	21.79	10.90

Table 4 EMMA: breakdown by subject (%).

Subtask	Anole	Anole-Zebra-CoT (Ours)
Overall	22.80	24.90
Scientific reasoning	30.33	32.79
Textbook question answering	36.08	29.75
Numeric commonsense	16.67	17.36
Arithmetic reasoning	15.58	18.98
Visual question answering	24.58	29.61
Geometry reasoning	20.50	23.01
Algebraic reasoning	25.27	24.56
Geometry problem solving	21.15	24.04
Math word problem	9.14	12.37
Logical reasoning	29.73	10.81
Figure question answering	24.54	28.25
Statistical reasoning	20.27	26.58

Table 5 MathVista: breakdown by subtask for base vs. our model (%).

Subtask	Anole	Anole-Zebra-CoT (Ours)
Overall	8.50	21.80
Quantitative reasoning	8.78	21.81
Spatial reasoning	8.23	22.08
Positional reasoning	8.82	19.85
Attribute reasoning	9.76	25.61
Stylistic reasoning	10.00	24.44
Other	5.56	18.52

Table 6 Visual Logic: breakdown by subtask (%).

Category	Anole	Anole-Zebra-CoT (Ours)
Overall	26.46	31.25
Art Style	19.66	35.04
Counting	19.17	15.00
Forensic detection	0.00	20.45
Functional correspondence	17.69	22.31
IQ test	26.67	23.33
Jigsaw	11.33	39.33
Multi-view reasoning	48.12	21.05
Object localization	50.82	45.90
Relative depth	38.71	41.94
Relative reflectance	29.10	27.61
Semantic correspondence	19.42	17.99
Spatial relation	41.26	57.34
Visual correspondence	21.51	26.16
Visual similarity	30.37	44.44

Table 7 Blink: breakdown by category (%).

D Impact Statement

All data sourced in this work were either publicly available under open licenses or generated synthetically. We ensured that all original content and assets used in the dataset creation process respect copyright and licensing terms. No human subjects were involved, and we do not foresee any direct harm to individuals or communities as a result of this work. The dataset is intended solely for academic research to improve multimodal reasoning capabilities in AI systems.

E Licenses

We list the licenses involved in this work as follows,

- Anole-7B model is under *Chameleon Research License*.
- ImageNet dataset is under *BSD 3 license*.
- Visual CoT dataset is licensed under *CC BY 4.0*
- MATH dataset ([Hendrycks et al., 2021](#)) is under *MIT License*.
- OpenStax Physics books are license under *CC BY 4.0*.
- MIT OCW Physics lecture notes under *CC BY 4.0*.
- Maze datasets is licensed under *CC BY 4.0*.