ARCTraj: A Dataset and Benchmark of Human Reasoning Trajectories for Abstract Problem Solving

Sejin Kim GIST
Hayan Choi VToV
Seokki Lee GIST
Sundong Kim* GIST

Abstract

We present ARCTraj, a large-scale dataset of human reasoning trajectories collected from interactive sessions on the Abstraction and Reasoning Corpus (ARC), a visual reasoning benchmark that challenges solvers to induce patterns from input-output grid pairs. While ARC provides only static examples, ARCTraj captures the whole sequence of high-level, object-centric actions humans take to solve these tasks, revealing intermediate steps typically hidden in conventional datasets. Collected via the O2ARC web interface, the dataset includes over 10,000 trajectories aligned with a Markov Decision Process (MDP) structure and enriched with metadata such as timestamps, user IDs, and success labels. ARCTraj has enabled diverse applications across reinforcement learning, sequence modeling, and generative planning, powering models such as PPO, World Models, Decision Transformers, GFlowNets, and diffusion agents. We further analyze the dataset to uncover behavioral patterns in spatial selection, color attribution, and strategy convergence. Together, these contributions position ARCTraj as a structured and interpretable resource for studying human-like reasoning and building cognitively informed learning systems.

1 Introduction

Understanding how humans reason and solve problems is a longstanding goal in artificial intelligence (AI). Human problem-solving often involves conceptual abstraction, attention shifts, and flexible strategy use,abilities that remain difficult for machines to emulate. The Abstraction and Reasoning Corpus (ARC) [7] was introduced to benchmark such capabilities through grid-based tasks where solvers must infer and apply rules from a small set of input-output examples. While ARC has inspired a wide range of approaches, including program synthesis [3], neuro-symbolic models [18], and test-time learning [2], it only provides static examples, making it difficult to analyze or model the dynamic reasoning processes humans use to solve these tasks.

To address this limitation, we present **ARCTraj**, a large-scale dataset of human reasoning trajectories collected while solving ARC tasks. Each trajectory captures a temporally ordered sequence of object-level actions (i.e., moving objects, rotating objects, and flipping objects) that transform an input grid into its correct output. These logs were collected through the O2ARC web interface [20], designed to support natural human interaction with ARC problems. Each trajectory is annotated with metadata such as timestamps, task identifiers, and success labels, enabling both learning and analysis.

Compared to existing human ARC datasets, ARCTraj offers several advantages. Whereas H-ARC [16] captures pixel-level edit logs and the ARC-Interactive-History-Dataset [22] records low-level cell and operation sequences, ARCTraj provides object-centric actions with consistent formatting across all 400 ARC training tasks. In addition, its public platform (O2ARC) supports visual inspection, replay, and future data expansion, making the dataset extensible and accessible.

*Corresponding author.

Preprint. Under review.

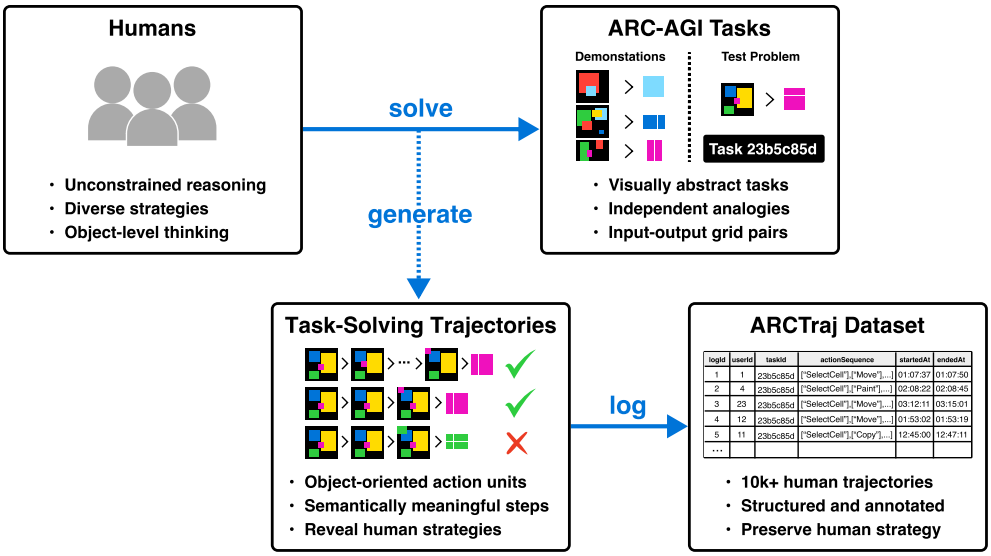


Figure 1: Overview of the ARCTraj data collection process. Users solve ARC tasks through the O2ARC platform by interacting with grid-based objects. Their actions are recorded step-by-step to form semantically rich, temporally ordered trajectories.

ARCTraj enables two main lines of research. First, it has been used to train learning agents in various paradigms, including reinforcement learning (e.g., PPO in ARCLE [14]), offline diffusion models [13], world models [15], and generative policies [10, 19]. Second, it allows analysis of human behavior in ARC tasks. In this work, we investigate spatial selection preferences, color source attribution, and strategy variation across users, highlighting the diversity and structure in human reasoning.

In summary, ARCTraj bridges a gap in the ARC ecosystem by providing dynamic, interpretable records of human reasoning. It supports both behavioral studies and cognitively inspired model development, making it a versatile resource for researchers interested in abstraction, planning, and generalization.

2 Related Work

Abstraction and Reasoning Corpus (ARC) The Abstraction and Reasoning Corpus (ARC) [7] is a benchmark to test human-like generalization in abstract reasoning tasks. Each ARC task comprises a few input-output grid pairs, requiring solvers to induce and apply conceptual transformations. It has motivated research across various paradigms, including program synthesis [3, 5, 6], neuro-symbolic reasoning [4, 18, 23], and test-time training [2, 9, 17], many of which were featured in the ARC Prize 2024 Technical Report [8]. However, ARC only provides input-output pairs, limiting insight into the dynamic reasoning steps that underlie human problem solving. In this work, we use the term ARC to refer specifically to the 400 training tasks from the ARC-AGI-1 benchmark, which serves as the basis for all trajectories collected in ARCTraj.

Human Trajectory Datasets for ARC Recent efforts have sought to capture human reasoning on ARC tasks through various forms of interaction data. LARC [1] collects natural language descriptions of task solutions, offering a semantic perspective but lacking action-level granularity. Fast and Flexible [11] and its successor H-ARC [16] describe the same pixel-level edit actions collected during the ARC task-solving dataset. While valuable, these low-level records are difficult to map onto structured reasoning steps or integrate into learning frameworks. The ARC-Interactive-History-Dataset recently introduced logs from the BrainGridGame (BGD) interface [22], capturing cell-level and operation-level actions over time. While this dataset supports MDP-like interpretations and includes many trajectories, it lacks standardized task coverage and object-level abstraction. It has not yet been integrated into downstream learning environments or model training pipelines.

3 The ARCTraj Dataset

ARCTraj is a large-scale dataset that captures human reasoning trajectories for solving abstract visual tasks from the ARC benchmark. Unlike the original ARC dataset, which only provides static input-output pairs, ARCTraj records temporally ordered, semantically grounded action sequences taken by humans while solving ARC tasks. This structure offers a high-resolution view of goal-directed behavior and supports research in sequential decision making, intention inference, and learning from demonstration.

3.1 Task and Collection Protocol

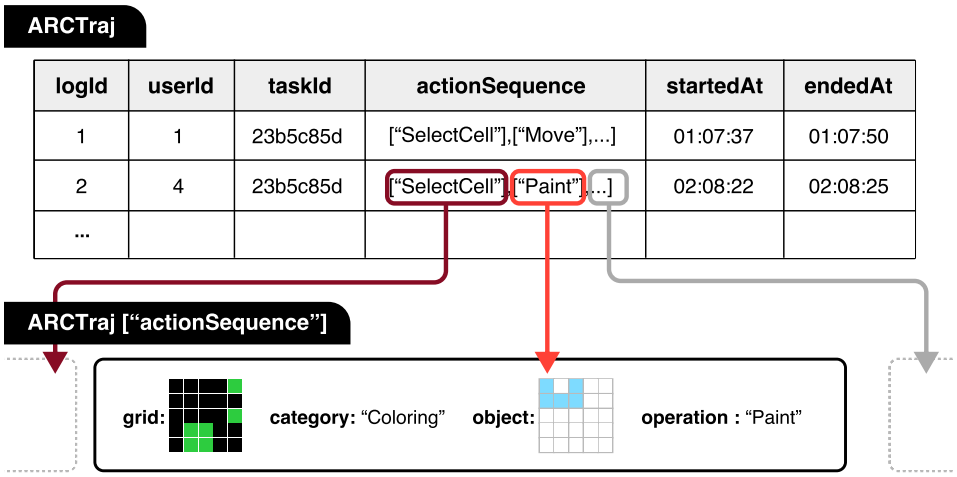


Figure 2: Example visualization of a single log in ARCTraj. Each action includes its category, operation, and associated grid and object state, forming a structured state-action unit.

All trajectories were collected using **O2ARC 3.0** [20], a custom web interface replicating the core mechanics of ARC tasks in an interactive, user-friendly format. Each task presents at least one input-output example and provides a manipulable input grid. Users interact with the grid through actions applied to *objects*, where each object is an automatically segmented group of adjacent colored pixels. Supported operations include `move`, `color`, `delete`, or `copy`, applied via clicks, keyboard shortcuts, or drag-and-drop. ARCTraj captures user actions at a conceptually coherent level aligned with human reasoning by operating on semantically meaningful objects rather than individual pixels.

Participants were instructed to solve tasks freely, without time limits or strategic constraints. All actions were automatically logged with timestamps and grid states. Each trajectory consists of alternating `selection` and `operation` steps, forming MDP-compatible state-action sequences suitable for downstream learning applications. In total, over 300 users contributed trajectories for all 400 training tasks from ARC-AGI-1, the official training split of the ARC benchmark.

3.2 Dataset Statistics

Table 1: Summary statistics of the ARCTraj dataset.

Metric	Value
Number of trajectories	10,672
Number of unique ARC tasks	400
Number of users	327
Mean trajectory length	9.8
Std. dev. of trajectory length	5.6
Success rate	82.3%
Most frequent action type	`move`
Average time per task	42.7 sec

The dataset includes 10,672 complete trajectories, averaging 9.8 actions per trajectory (std: 5.6). Each trajectory includes metadata such as task ID, user ID, timestamps, action types, and success labels. Participants solved various ARC tasks involving symmetry, object grouping, spatial transformations, and numeric rules. Most actions were completed within 43 seconds on average, indicating natural problem-solving pacing. The `move` operation reflects the design's emphasis on object-level interactions that match human intuitions for manipulating abstract visual elements.

3.3 Comparison with Existing ARC Datasets

Several prior datasets have attempted to capture human reasoning on ARC tasks through logs of natural language, pixel-level interactions, or interface actions. **Fast and Flexible** [11] and its follow-up study **H-ARC** [16] describe the same dataset of pixel-level edit sequences collected from a custom ARC-solving interface. These datasets provide fine-grained logs of human actions at the pixel level and have been used to study planning and temporal patterns. However, the recorded actions are low-level and lack object abstraction, making it difficult to model behavior using structured learning frameworks such as reinforcement learning or sequence modeling.

More recently, the **ARC-Interactive-History-Dataset** [21], collected through the BrainGridGame interface [22], logs cell-level and operation-level actions as humans solve ARC tasks. While BGD provides structured logs with semantic labels and supports MDP-compatible interpretations, it differs from ARCTraj in several important ways: (i) the task set is not aligned with the official ARC benchmark (e.g., ARC-AGI-1), (ii) object-level operations are inferred rather than explicitly logged, and (iii) no public platform exists for interaction replay or exploration.

In contrast, **ARCTraj** offers a task-aligned, object-centric, and publicly explorable dataset built on the official ARC training split. It features structured state-action sequences with clearly defined object boundaries and operations, captured through a uniform user interface. Moreover, the O2ARC platform [20] enables replay, inspection, and visualization of each trajectory, facilitating both reproducibility and in-depth behavioral analysis.

Overall, ARCTraj complements existing efforts by combining the semantic structure of object-level actions with compatibility for downstream modeling, while providing broader task coverage and an accessible ecosystem for interactive analysis.

4 Human Trajectory Analyses

ARCTraj is not only a dataset of recorded behavior but also a foundation for analyzing strategic diversity and cognitive patterns in human problem-solving. This section presents a series of structured analyses on human trajectories, aiming to reveal both low-level interaction biases and high-level reasoning dynamics embedded in the problem-solving process. We use various methods, including statistical summaries, trajectory clustering, and visualization, to extract insights that may inform the design of human-aligned models.

We organize our analyses around three core research questions (RQs) spanning micro-level behaviors and macro-level reasoning structures. These questions cover a broad spectrum of cognitive aspects, from where humans focus their attention and how they choose colors, to the diversity of problem-solving strategies and the semantic coverage of existing symbolic abstractions.

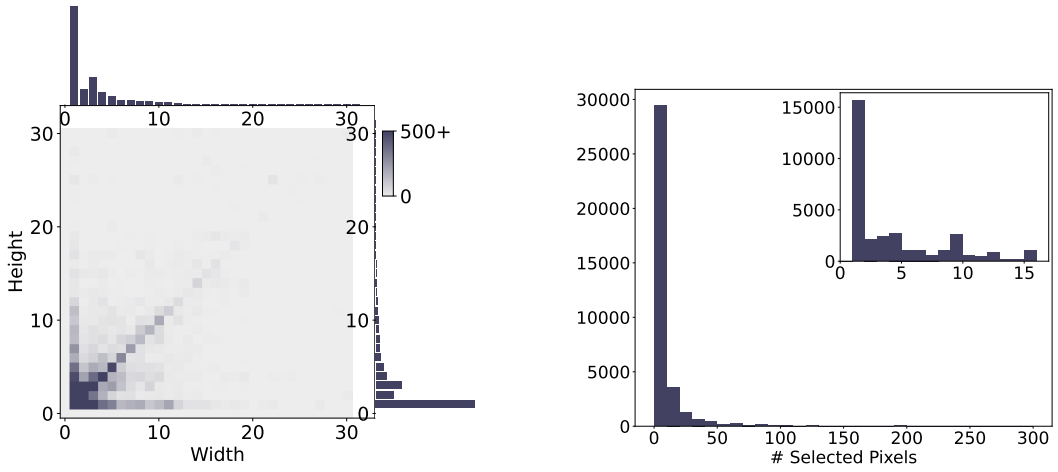
RQ 1. **Do humans exhibit spatial or object-level selection biases when solving ARC tasks?** This question investigates action tendencies such as preferred regions of interaction, the relationship between the number of objects and trajectory length, and whether particular objects are disproportionately selected.

RQ 2. **Where do the colors used in test-time outputs originate, and how do they relate to human color selection strategies?** We examine whether color choices are derived from input grids, reference outputs, or latent reasoning, thereby informing the design of inductive biases in generative models.

RQ 3. **What strategic patterns and clustering structures emerge across human trajectories solving the same ARC task?** By analyzing variation across solution paths, we uncover how humans approach the same goal differently and how such variations relate to transformation patterns or inferred intentions.

4.1 Biases in Human Grid Selections

To address RQ1, we examine whether humans exhibit systematic selection biases when interacting with ARC grids. In ARCTraj, each *selection* action may correspond to a single pixel, a user-specified region, or an object-level selection that implicitly includes multiple pixels. To unify these heterogeneous selection types, we compute the *bounding box* for each selection—the smallest axis-aligned rectangle that fully contains all selected pixels. We then analyze the distribution of these bounding boxes in terms of their height, width, and area, enabling a consistent characterization of selection scale and shape across the dataset.



(a) Joint distribution of selection height and width.

(b) Histogram of the number of selected pixels.

Figure 3: Distributions of human selection behavior in ARC tasks. Left: Selections are concentrated in compact shapes such as 1×1 to 3×3, with square and bar-shaped regions dominating. Right: Most selections cover fewer than 20 pixels, supporting the preference for local and perceptually salient regions.

As shown in Fig. 3, we identify three dominant tendencies in human selections: (i) selected areas are predominantly small (typically less than 3×3), (ii) selections are often square-shaped ($n \times n$), and (iii) bar-shaped selections ($n \times 1$ or $1 \times m$) also frequently occur. These findings suggest a general preference for local reasoning and perceptual regularity in human problem solving. The left panel of Fig. 3 shows the joint distribution of selection height and width across all selections, revealing an intense concentration in the 1×1 to 3×3 range. A diagonal ridge indicates a square-shape bias, while off-diagonal clusters correspond to frequent horizontal or vertical bar-shaped selections. The marginal histograms reveal firm peaks at width $= 1$ and height $= 1$, indicating a bias toward compact and axis-aligned regions. Complementing this, the right panel shows the distribution of the number of selected pixels per action. Most selections involve no more than 16 pixels. Interestingly, local peaks appear near square numbers (e.g., 1, 4, 9, 16), likely reflecting the high frequency of $n \times n$ square-shaped selections observed in the left panel. This reinforces that humans focus on perceptually salient, compact regions, regardless of task complexity.

Research Direction 1(a): Temporal Dynamics of Selection Behavior. Future work could investigate how selection size and shape evolve during problem-solving. Do humans start with small exploratory selections and later switch to larger ones once a transformation pattern is identified? Or do they first examine the global structure before zooming into specific details? Temporal analysis of selection sequences could provide insight into human attention shifts and inform AI agents' curriculum or phase-based learning strategies.

Research Direction 1(b): Perceptual Features and Selection Probability. Another promising avenue is to examine whether features such as color contrast, spatial isolation, or proximity to grid boundaries influence the likelihood of selection. This could be tested via controlled manipulation of object arrangements and saliency. Modeling this relationship may lead to predictive models of human attention or selection likelihood, which could be incorporated into attention-guided architectures or human-AI collaborative systems.

4.2 Color Source Attribution in Test Outputs

To address RQ2, we examine the origins of colors used in the test output grid and how they relate to human color selection strategies. Color is one of the most critical factors when solving ARC tasks, yet understanding how humans select colors presents unique challenges in our dataset collection methodology.

Analyzing both the task and the collected trajectory reveals that test output colors typically originate from limited sources. Among 400 ARC training tasks, 266 have solutions where colors could be selected exclusively from the color set of the test input grid, while 134 tasks require colors from the union of the color set of the test input grid and the example outputs grid. Interestingly, we found no cases where colors unique to example inputs were necessary for correct solutions, even when considering the complete set of potential sources (test input + example output + example input). This pattern suggests a deliberate task design constraint that limits the search space for potential color sources, focusing primarily on test inputs and secondarily on example outputs while avoiding reliance on example inputs for solution colors.

While our trajectory data does not explicitly record where users sourced their colors, statistical color selection patterns align closely with these potential sources. Users consistently select colors in the test input or example outputs, even without explicit color sampling tools. This suggests humans perform implicit source attribution when reasoning about color transformations, mentally tracking color origins and relationships across different grid examples.

Table 2: Distribution of color sources in ARC tasks. For 66.5% of tasks, all required colors appear in the test input grid. The rest require colors from both the test input and example output grids. No task requires colors exclusive to the example input.

Color Source	**# of Tasks**	**%**
Test Input	266	66.5
Test Input + Example Output	134	33.5
Example Input Only	0	0.0
Others	0	0.0

Research Direction 2(a): Trajectory Logging with Color Origin Capture DSL. Future research would benefit from developing more sophisticated trajectory recording interfaces incorporating explicit color origin tracking mechanisms. By extending current DSLs with formal color sourcing operators (e.g., `sample_color(grid, x, y)` or `apply_color_transformation(rule)`), researchers could create interfaces that allow users to directly sample colors from different grids while precisely logging these conceptual connections. Such integrated systems would document which colors were selected and their relational origins, capturing the mental models humans construct when reasoning about color transformations. These enhanced trajectory capture tools would generate richer datasets that more accurately reflect how humans establish color correspondences across examples, enabling more precise evaluation of computational models against human color selection strategies. Implementing and testing these extended DSLs could significantly advance our understanding of the cognitive processes underlying abstract visual reasoning tasks.

Research Direction 2(b): Generalized Origin Tracking Across Multiple Elements. Building on insights from color origin analysis, future research should explore how humans source and transfer various elements beyond colors when solving ARC tasks. As colors may originate from test inputs or example outputs, other critical problem elements—such as object selection patterns, grid dimensions, transformation sequences, and spatial relationships—likely derive from specific examples within the task. Researchers could develop comprehensive origin tracking frameworks that identify how humans extract and repurpose information across multiple dimensions of the problem space. For instance, when users create specific grid structures, do they primarily reference example output configurations? When selecting objects of particular sizes, are they influenced more by test inputs or example patterns? This multi-dimensional origin analysis would better understand how humans perform cross-example analogical reasoning, revealing which task elements are anchors for different solution development aspects. Such research could significantly advance our understanding of the hierarchical and relational nature of human abstract reasoning, potentially informing more sophisticated computational models that similarly draw information from appropriate sources when building solutions.

4.3 Shared Intentions and Strategy Patterns

To address RQ3, we investigate whether humans solving the same ARC task exhibit converging strategies or follow diverse solution paths. Rather than comparing complete trajectories, we focus on mid-sequence decisions that reflect shared *intentions*, concrete choices about *what* region to act on and *how* to act upon it.

In ARCTraj, each trajectory comprises alternating focus and transformation steps. A focus action highlights a rectangular grid region, which may correspond to an entire object, a shape fragment, or a structured area. After one or more such steps, the user applies a transformation—such as moving, coloring, deleting, or copying—to the focused region. Human solvers do not strictly alternate between focusing and acting; they often examine several areas of succession to inspect the grid or compare subgoals before committing to a transformation. For example, a user might highlight multiple red squares before recoloring a blue one to match, reflecting a search-and-align strategy.

Table 3: Distribution of selection actions preceding each operation. Most operations follow 1–4 selections, suggesting localized exploration before committing to a decision.

Length	Count	%	Cum. %
1	23,632	63.7	63.7
2	5,451	14.7	78.4
3	3,343	9.0	87.4
4	1,379	3.7	91.1
⋮	⋮	⋮	⋮
386	1	0.0	100.0

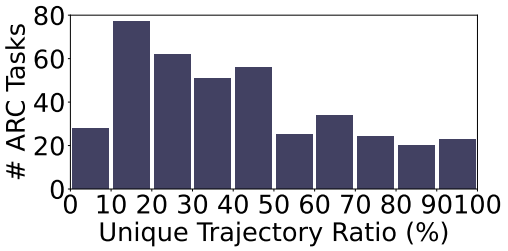


Figure 4: Histogram of task-level trajectory uniqueness. Tasks with low uniqueness (left) exhibit high convergence in human strategies. High uniqueness (right) indicates diverse or ambiguous solution paths.

Our analysis reveals that 63.7% of operations are preceded by a single selection, and over 90% occur within four selections (Table 3). This suggests that a few attentional shifts typically drive human planning before executing a concrete transformation. The short span between selections and operations indicates that humans often perform quick exploratory lookups before committing to a goal-directed action.

To formalize mid-level convergence, we define a shared *intention* as pairing a selection region and an operation type that recurs across different users solving the same task. In other words, if two users both select a 2×2 red square in the lower-left corner and change it to blue, this counts as an instance of shared intention, even if the rest of their actions diverge. This notion captures agreement on *what to do* at a specific point in the problem-solving process, without requiring their entire trajectories to align.

To operationalize this, we extract all (selection, operation) pairs from each trajectory and cluster them within each task based on spatial and semantic similarity. Some tasks exhibit strong convergence, where most users perform the same key transformation on the same grid region. Others show high divergence, with users selecting different substructures or applying varied operations. As shown in Figure 4, tasks with low trajectory uniqueness often correspond to visually salient or intuitively structured solutions. In contrast, ARC tasks with high uniqueness are more ambiguous or allow for multiple valid approaches.

Research Direction 3(a): Strategy Grammar and Diversity Mapping. Future work could formalize intention clusters into a compositional strategy grammar that captures reusable abstraction templates across tasks. Identifying common strategy motifs (e.g., "color and duplicate," "fold and align") would enable interpretable models and human-aligned planning systems.

Research Direction 3(b): Intention Prediction and Curriculum Design. Another promising direction is to train models that predict the distribution of human intentions based on task features. This could inform the sequencing of curriculum tasks, scaffold learning from easier to more diverse cases, and support adaptive tutoring systems that anticipate user strategies and provide personalized guidance.

5 Learning with ARCTraj

ARCTraj enables various downstream learning applications by providing structured, object-level, and temporally ordered human trajectories that can be interpreted as state-action sequences. This structure supports both reinforcement learning (RL) and sequence modeling approaches, offering rich supervision derived from human reasoning. This section describes how ARCTraj has been used in interactive and non-interactive learning settings and summarizes empirical results from prior work.

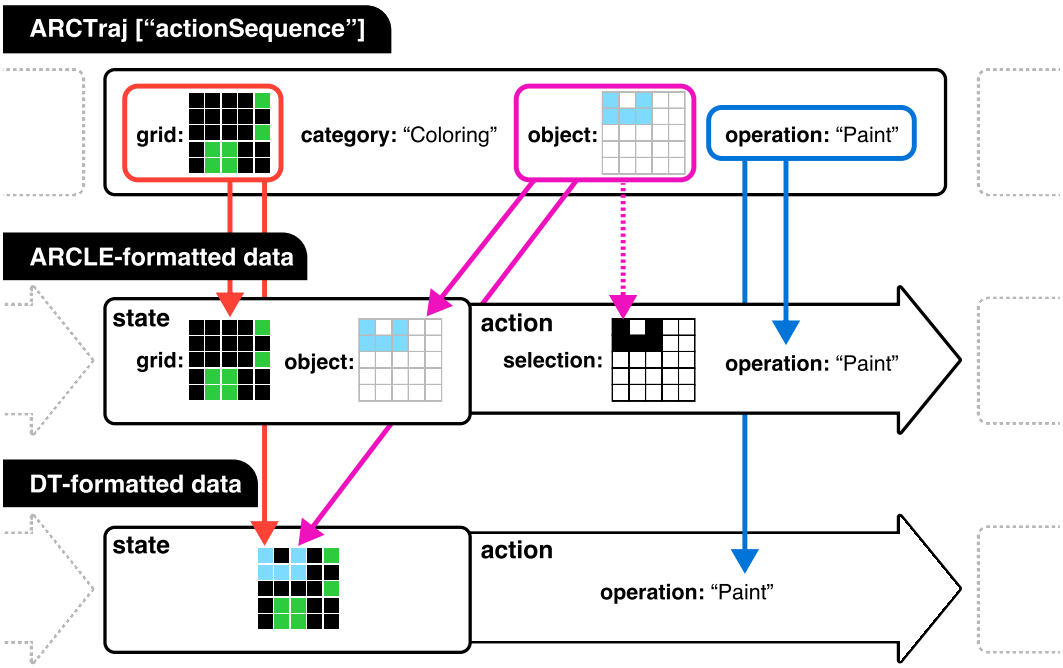


Figure 5: Preprocessing ARCTraj for downstream learning. For RL environments such as ARCLE, ARCTraj is filtered to retain only `operation` actions and is mapped to a Markovian state-action format using `grid`, `object`, and `operation`. For sequence models such as Decision Transformer, only `grid` and `operation` are used, omitting intermediate objects and environment interaction.

5.1 Offline Reinforcement Learning with ARCTraj

ARCTraj can be naturally interpreted as a sequence of (s_t, a_t) pairs, where each state s_t encodes the grid and active object, and each action a_t denotes a high-level transformation applied by a human solver. This structure directly supports offline reinforcement learning (RL) and has been used to construct ARCLE [14], a custom environment aligned with ARCTraj's interaction interface. In ARCLE, trajectories are filtered to retain transformation steps while discarding intermediate selections, enabling agents to learn symbolic manipulation policies grounded in human decisions. PPO-based agents trained in this setting demonstrate the feasibility of learning task-generalizable behaviors from human data.

Several generative planning agents also leverage ARCTraj. LDCQ [13] augments the trajectories by interpolating intermediate grid states between human actions, enabling diffusion models to synthesize coherent, multi-step plans. DreamerV3-based models [15] train world models on ARCTraj to capture latent dynamics and perform analogical generalization across structurally similar tasks. GFlowNet-based methods [10] treat human trajectories as samples from an implicit distribution over valid solutions, training agents to generate diverse, goal-consistent behaviors that mirror the variation seen in human problem solving.

Together, these applications show that ARCTraj supports both imitation-style policy learning and generative behavior modeling. Its structured and semantically grounded trajectories enable agents to reason symbolically, generalize across tasks, and sample solutions in a manner consistent with human strategies.

5.2 Trajectory-Based Learning Without Explicit Environments

ARCTraj also supports learning in non-interactive frameworks without requiring simulation environments. As shown on the right of Fig. 5, trajectories can be simplified by extracting sequences of `grid` states and `operation` actions, omitting the object-level details. This compact format is well-suited for sequence models such as Decision Transformers [19], which treat trajectory length and success as proxy rewards and learn policies directly from demonstrations.

In addition, intention-based approaches further utilize ARCTraj to align low-level actions with latent subgoals. For instance, recent work [12] introduces intention-conditioned supervision to enhance generalization and interpretability. These studies show that incorporating inductive biases, such as latent goal structure, can improve trajectory modeling even in the absence of explicit environments.

5.3 Summary of Learning Outcomes

Table 4 summarizes key methods and findings across interactive and non-interactive paradigms using ARCTraj. These applications demonstrate that ARCTraj is a flexible foundation for training policies, planning agents, and generative models. Its structured and interpretable design supports both symbolic and neural architectures, facilitating comparisons across modeling approaches and enabling progress on problems such as alignment, abstraction, and data-efficient learning.

Table 4: Summary of representative methods and findings from prior work leveraging ARCTraj.

Research	**Setting**	**Model**	**Key Findings**
ARCLE [14]	Online RL	PPO	Demonstrated the feasibility of training in an ARCTraj-compatible MDP environment.
LDCQ [13]	Offline RL	Diffusion	Augmented ARCTraj with intermediate states to enable generative plan synthesis.
DreamerV3 [15]	Offline RL	World Model	Enabled analogical generalization by modeling latent dynamics transferable across tasks.
GFlowNet [10]	Offline RL	GFlowNet	Sampled goal-directed trajectories aligned with expert-like solutions.
Decision Transformer [19]	Offline	Transformer	Learned trajectory-conditioned policies from human demonstrations.
Intention Learning [12]	Offline	Transformer	Enhanced generalization via subgoal alignment and intention-conditioned modeling.

6 Conclusion

ARCTraj captures over 10,000 human trajectories on ARC tasks, offering fine-grained, step-by-step records of people engaging with abstract visual reasoning problems. Unlike conventional datasets that only provide static input-output pairs, ARCTraj logs temporally ordered, object-level actions grounded in human perceptual and symbolic understanding. This structure provides a window into intermediate reasoning processes and enables detailed modeling of how people plan, adapt, and transform problem representations during task solving.

These structured trajectories have already supported a wide range of learning settings. They enable reinforcement learning agents to be trained from human demonstrations, guide generative planners through trajectory-based data augmentation, and support intention-aware modeling that uncovers latent subgoals. Moreover, the dataset has proven compatible with a variety of architectures—including PPO, diffusion models, GFlowNets, and decision transformers—highlighting its versatility across paradigms.

Beyond supporting modeling, ARCTraj enables the analysis of behavioral and cognitive patterns that are difficult to observe in final outputs alone. Our findings reveal consistent regularities in selection behavior, biases in color attribution, and clustered transformation strategies, which offer a new lens on human abstract reasoning and task decomposition. Such insights can inform the design of inductive biases, curriculum structures, and evaluation protocols for cognitively aligned AI.

References

[1] Samuel Acquaviva, Yewen Pu, Marta Kryven, Theodoros Sechopoulos, Catherine Wong, Gabrielle Ecanow, Maxwell Nye, Michael Tessler, and Joshua B. Tenenbaum. Communicating Natural Programs to Humans and Machines. In *NeurIPS Datasets and Benchmarks*, 2022.

[2] Ekin Akyürek, Mehul Damani, Linlu Qiu, Han Guo, Yoon Kim, and Jacob Andreas. The Surprising Effectiveness of Test-Time Training for Abstract Reasoning. *arXiv:2411.07279*, 2024.

[3] Shraddha Barke, Emmanuel Anaya Gonzalez, Saketh Ram Kasibatla, Taylor Berg-Kirkpatrick, and Nadia Polikarpova. HYSYNTH: Context-Free LLM Approximation for Guiding Program Synthesis. *arXiv:2405.15880*, 2024.

[4] Paweł Batorski, Jannik Brinkmann, and Paul Swoboda. NSA: Neuro-Symbolic ARC Challenge. *arXiv:2501.04424*, 2025.

[5] Mikel Bober-Irizar and Soumya Banerjee. Neural Networks for Abstraction and Reasoning. *Scientific Reports*, 14(1):27823, 2024.

[6] Natasha Butt, Blazej Manczak, Auke Wiggers, Corrado Rainone, David Zhang, Michaël Defferrard, and Taco Cohen. CodeIt: Self-Improving Language Models with Prioritized Hindsight Replay. *arXiv:2402.04858*, 2024.

[7] François Chollet. On the Measure of Intelligence. *arXiv:1911.01547*, 2019.

[8] Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. Arc Prize 2024: Technical Report. *arXiv:2412.04604*, 2024.

[9] Daniel Franzen, Jan Disselhoff, and David Hartmann. The LLM ARChitect: Solving ARC-AGI Is A Matter of Perspective, 2024.

[10] Sanha Hwang, Sejin Kim, Seungpil Lee, and Sundong Kim. Solution Augmentation for ARC-AGI Problems Using GFlowNet: A Probabilistic Exploration Approach. *Transactions on Machine Learning Research (submitted)*, 2024.

[11] Aysja Johnson, Wai Keen Vong, Brenden M. Lake, and Todd M. Gureckis. Fast and Flexible: Human Program Induction in Abstract Reasoning Tasks. In *CogSci*, 2021.

[12] Sejin Kim, Hosung Lee, and Sundong Kim. Addressing and Visualizing Misalignments in Human Task-Solving Trajectories. In *ICLR Workshop on Bidirectional Human-AI Alignment*, 2025.

[13] Yunho Kim, Jaehyun Park, Heejun Kim, Sejin Kim, Byung-Jun Lee, and Sundong Kim. Diffusion-Based Offline RL for Improved Decision-Making in Augmented ARC Task. *arXiv preprint arXiv:2410.11324*, 2024.

[14] Hosung Lee, Sejin Kim, Seungpil Lee, Sanha Hwang, Jihwan Lee, Byung-Jun Lee, and Sundong Kim. ARCLE: The Abstraction and Reasoning Corpus Learning Environment for Reinforcement Learning. In *CoLLAs*, 2024.

[15] Jihwan Lee, Woochang Sim, Sejin Kim, and Sundong Kim. Enhancing Analogical Reasoning in the Abstraction and Reasoning Corpus via Model-Based RL. In *IJCAI Workshop on Interactions between Analogical Reasoning and Machine Learning*, 2024.

[16] Solim LeGris, Wai Keen Vong, Brenden M Lake, and Todd M Gureckis. H-ARC: A Robust Estimate of Human Performance on the Abstraction and Reasoning Corpus Benchmark. *arXiv:2409.01374*, 2024.

[17] Wen-Ding Li, Keya Hu, Carter Larsen, Yuqing Wu, Simon Alford, Caleb Woo, Spencer M Dunn, Hao Tang, Michelangelo Naim, Dat Nguyen, Wei-Long Zheng, Zenna Tavares, Yewen Pu, and Kevin Ellis. Combining induction and transduction for abstract reasoning. In *ICLR*, 2025.

[18] Isaac Liao and Albert Gu. CompressARC: ARC Solving Without Data Augmentation or Pretraining, 2024.

[19] Jaehyun Park, Jaegyun Im, Sanha Hwang, Mintaek Lim, Sabina Ualibekova, Sejin Kim, and Sundong Kim. Unraveling the ARC Puzzle: Mimicking Human Solutions with Object-Centric Decision Transformer. In *ICML Workshop on Interactive Learning with Implicit Human Feedback*, 2023.

[20] Suyeon Shim, Dohyun Ko, Hosung Lee, Seokki Lee, Doyoon Song, Sanha Hwang, Sejin Kim, and Sundong Kim. O2ARC 3.0: A Platform for Solving and Creating ARC Tasks. In *IJCAI Demo*, 2024.

[21] Simon Strandgaard. ARC-Interactive-History-Dataset, 2024.

[22] Simon Strandgaard. Brain Grid Game, 2024.

[23] Yudong Xu, Elias B. Khalil, and Scott Sanner. Graphs, Constraints, and Search for the Abstraction and Reasoning Corpus. In *AAAI*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction describe the main contributions of the paper accurately, including the collection of 13,000+ object-centric human reasoning trajectories for ARC tasks, and their use in various downstream learning paradigms. See Abstract and Sec. 1.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sec. 4 discusses limitations such as lack of explicit color origin tracking and varying strategic diversity, and proposes future research directions to address these.

   Justification: Sec. 4 discusses several limitations, including the lack of explicit color origin tracking (Sec. 4.2) and the challenges in clustering diverse human strategies (Sec. 4.3). The authors acknowledge these limitations and propose directions for future work to address them, such as extending the trajectory logging interface and developing more structured strategy grammars.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include any theoretical results or formal proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The data collection process (Fig. 1), dataset structure (Fig. 2), and preprocessing for learning ARC tasks (Fig. 5) are described in detail. Also, Sec. 4 includes three research questions (RQs) and their own analysis results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: ARCTraj is intended as a public benchmark dataset. The data is already uploaded to HuggingFace (`https://huggingface.co/datasets/SejinKimm/ARCTraj`, and final instructions for access will be included in the camera-ready version.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: While the paper itself does not conduct new model training, it summarizes how ARCTraj has been used in prior studies and includes descriptions of model types, formats, and data preprocessing used in learning applications (Sec. 5).

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: While the paper focuses on dataset analysis rather than model comparison, it reports distributions, variances, and cumulative statistics (e.g., Fig. 3, Table 2, Table 3, and Fig. 4), which are sufficient for interpreting human behavior trends.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper does not include computational details for the downstream model training mentioned (e.g., PPO, GFlowNet). These are referenced from prior work without reporting computational resources.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The dataset was collected via voluntary user participation on a public interface with anonymized logs. No personally identifiable information is included.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Sec. 6 discusses the positive impact of aligning AI systems with human conceptual reasoning. No explicit discussion of potential misuse is included.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The ARCTraj dataset poses minimal risk for misuse. It does not include generative models or scraped content.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites all relevant prior datasets (e.g., ARC, H-ARC, LARC) and tools (e.g., BrainGridGame) with references.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: ARCTraj is introduced as a new asset and is documented extensively in Sec. 3, including format (Fig. 2), statistics (Table 1), and collection protocol.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Participants voluntarily generated trajectory data through the O2ARC web interface without direct monetary compensation. As described in O2ARC 3.0 [20], the platform incorporates several gamified features that encourage high-quality user engagement, such as task-level scoring, real-time rank display, and monthly leaderboard rankings. These mechanisms were effective in motivating participants to contribute valuable trajectories. The full instructions provided to users will be included in the supplemental material.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: The ARCTraj dataset contains only anonymized interaction logs of users solving visual reasoning tasks via the O2ARC platform. No personally identifiable information (PII), demographic data, or free-form text was collected. The recorded data consists solely of de-identified grid-based action sequences (e.g., object selection, movement, transformation), which cannot be linked to any individual. As the study did not involve sensitive data and posed no foreseeable psychological, physical, or legal risks, IRB approval was not required under our institution's guidelines.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [NA]

    Justification: LLMs were used only for minor writing assistance, such as grammar correction and formatting, without influencing any methodological, analytical, or experimental aspect of the research.