Strategy Discovery for Long-Horizon Physical Manipulation Puzzles

Svetlana Levit (levit@tu-berlin.de)[1,5], Erfan Baradarantohidi[1,2,5], Jinan N. Allan[2,3,5], Thorsten Pachur[2,4,5] & Marc Toussaint[1,5]

[1]Technical University of Berlin, Germany
[2]Max Planck Institute for Human Development, Germany
[3]Clemson University, USA
[4]Technical University of Munich, Germany
[5]Science of Intelligence Cluster of Excellence, Germany

Abstract

To support the systematic study of long-horizon human strategies in physical problem solving, we introduce a set of 28 virtual physics puzzles that require sequential manipulation and involve diverse interaction mechanics such as pick-and-place, pushing, and tool use and creation. The paradigm provides rich interaction logs and object-state trajectories, enabling computational analyses of strategies. We analyze data from 38 participants and estimate puzzle difficulty and participant performance. Using trajectory-based clustering, we identify recurring solution strategies and quantify solution diversity and repeatability between attempts. The results show that puzzles requiring both tool creation and pushing are the most challenging to solve with the longest time-to-solution and the highest failure rates, as they combine the large search spaces of tool use with the challenges of precise control and less predictable dynamics. Notably, participants demonstrate strong improvement after solving the puzzles once, suggesting learning and reuse of discovered strategies.

Keywords: Sequential Manipulation; Strategies; Physical Reasoning; Tool Use

Introduction

Humans are highly effective at physical problem solving through interaction, trial-and-error, and strategy refinement. However, many physical reasoning problems cannot be solved in a single action: they require **long-horizon sequences of interdependent manipulations**, often under uncertainty about dynamics and control. Prior research on intuitive physics and physical reasoning has largely emphasized one-shot predictions or single-action solutions, while fewer paradigms capture multi-step physical problem solving with strategy changes and repetition. We address this gap by introducing a set of **28 virtual physics puzzles** for studying physical reasoning in sequential and dynamic physical problems: problems where success depends on selecting a plan, executing a series of manipulations, and adapting when dynamics make outcomes uncertain. To achieve the goal, participants must manipulate objects using diverse mechanics such as pick-and-place, pushing, and tool use and creation.

Our work addresses three main questions:

1. **Difficulty:** Which physical puzzles are more difficult and what features make them challenging?
2. **Strategies:** Do current puzzles support the discovery of distinct strategies? Which puzzles require strategy switching (changes at the parameter vs. plan level)?
3. **Exploration and repetitions:** How are exploration and retries linked to solution performance?

The virtual setting makes it possible to study physical reasoning in many diverse environments (e.g., ramps, narrow passages, inaccessible zones). It also supports the efficient collection of interaction and trajectory data to quantify not only success and solution time but also exploration, solution similarity, and reproducibility across attempts. Finally, our problems combine discrete and continuous planning decisions (e.g., what to use, in what order, and how to grasp), supporting comparisons between human solutions and artificial solvers (i.e., robotic planners) in challenging settings where both planning and execution matter.

Related Work

General physical problem solving requires a causal understanding of how physics in the world works and the ability to make inferences, learn, and plan, to solve novel problems. Early work examined trial-and-error and insight in animal tool-use (Köhler, 1925; Thorndike, 1898). One of the classic works on human problem solving (Newell & Simon, 1972) introduced a formal analysis method, the means-end analysis, which identified the essential features of a structured solution search. Dörner and Funke (2017) then expanded it from logic puzzles to less well-defined and more complex problems.

More recent research focuses on reasoning and planning in novel situations and the ability to build and use tools (Allen, Smith, & Tenenbaum, 2020; Mitko & Fischer, 2020).

Physical reasoning has been studied in simulation-based benchmarks such as PHYRE (Bakhtin et al., 2019) and Tool-Games/VirtualTools (Allen, Smith, & Tenenbaum, 2020), which support controlled comparisons between humans and computational models and highlight the role of rich scene understanding. However, these works primarily modeled single-action decisions.

Work on sequential physical problem solving addresses this shortcoming by emphasizing the role of multi-step planning and tool use for physical reasoning tasks (Toussaint et al., 2018; Yildirim et al., 2017).

At the intersection of robotics and cognitive science, several approaches compare and translate human physical problem solving strategies to synthetic or artificial solvers and robots (Lozano-Pérez & Kaelbling, 2014; Toussaint et al., 2018). Many robot manipulation and reinforcement learning (RL)

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

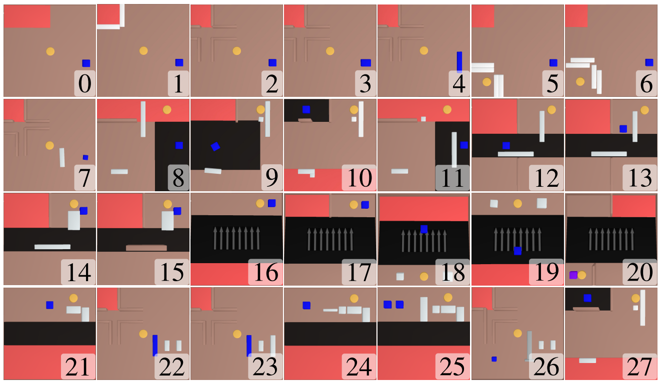


Figure 1: Puzzles 0-27. The agent (yellow) is controlled by a mouse to move the blue object(s) into the red goal area. Movable obstacles are white, walls are brown. Black areas are inaccessible to the agent. The grey arrows indicate a ramp.

benchmarks focus on reasoning about object dynamics, which is a very challenging part of physical reasoning, but only a few support tool use or tool creation. In Table 1, we list recent frameworks allowing human interaction.

Unlike previous studies that had to physically create environments (Karimpur et al., 2024; Toussaint et al., 2018), which required time and limited the number of problems a participant could explore, our virtual paradigm allows researchers to consider multiple different environments, which can include obstacles, inaccessible zones, ramps, and involve tool creation.

Virtual Physical Puzzles Experiment

In the newly designed tasks (Fig. 1), the participant uses the computer mouse to control the puzzle agent (the yellow circle) to manipulate other objects in the scene. The goal is always to move the blue *goal object(s)* into the red goal area. The agent can only move in 2D along the floor plane, but the objects can be moved in 3D (e.g. sliding down a ramp). Our interface allows for two types of manipulation: *Pick-and-Place*, where the agent grasps (attaches) an object in a fixed manner with a mouse click, and *Push*, where the agent pushes or kicks an object. All objects can be combined (glued) to build new tools, which can be manipulated in the same way. To simulate physics in our puzzles, such as applied forces in pushing, gravity, and friction, we used `Bullet` Physics SDK (Coumans, 2015).

We grouped puzzles by the dominant manipulation type:

1. **Pick-and-place (PnP)**: Push action is disabled. The control is more predictable, as contacts do not trigger interactions unless an action is executed by a mouse click. These puzzles increase in complexity by including obstacles, narrow passages, or interdependencies in moving and picking actions.
2. **Glue**: The puzzle solution requires objects to be combined (e.g., glued together). One of the challenges in gluing is positioning objects correctly, as the glued parts need to be close enough to each other and to the agent.
3. **Dynamic (Dyn)**: Puzzles involve reasoning about Newtonian dynamics (e.g., pushing an object up a ramp).
4. **Push**: The environment allows for pushing objects, and tool creation/use are essential.

Some puzzles combine multiple challenging elements, adding to the diversity of our environments (see Table 2). We define *tool-use* as using one object to push another, for instance, the goal object. Puzzles which can be solved by using available objects as tools we categorize as *tool-use*, and puzzles requiring creating new tools by gluing objects as *tool-build*. While tool-use is challenging due to the fine control needed, we found that creating a tool adds combinatorial complexity to the task. Importantly, even though for the Glue puzzles we expect participants to connect objects, we do not consider them as tool use, because they do not allow Push actions.

Participants

Data collection took place between July 2022 and July 2023 at the Max Planck Institute for Human Development in Berlin. Two initial phases with 11 participants included an interview to collect feedback on the usability of the setup. In the final two phases, 38 participants completed the experiment. The mean age was 25.26 (SD 4.4); 14 participants were men and 24 women. For all further analyses, we rely on the data from the final phases ($N = 38$).

Procedure

Participants were presented with a series of 28 computer-based physical puzzles in a fixed order (Fig. 1). They were able to restart or skip a puzzle at any point. The experiment included a tutorial, two runs of the puzzle series, and a follow-up survey.

Tutorial: The tutorial explained the general goal of the puzzles, as well as how to use the computer controls to manipulate objects: learning to pick, push, and glue and unglue objects.

Runs: After attempting all puzzles (1st run), participants were asked to solve them again in the same order (2nd run).

Difficulty Rating: In the second run, participants could rate the difficulty of each puzzle from 1 (easy) to 5 (hard).

Follow-Up Survey: After completing the experiment, participants were directed to a survey in which they provided basic demographic information (e.g., age, gender, level of education), as well as their daily experience with computers.

Participants who completed the full trial were compensated equally, regardless of their performance on the tasks and their time spent, or received a partial payment if they chose to stop earlier. The expected duration was 60–90 minutes.

Data Collection

In the virtual physics puzzles program, we recorded complete mouse movement data, the interaction events, the intermediate states of the puzzles, and meta-information: start and end time, objects interacted with, solved flag. For each solution attempt, we saved two files: *interaction* and *frame* data. Interaction data contain the mouse position, type and time of interaction, and object names. Interactions are stored only when an interaction

Table 1: Interactive Frameworks for Physical Reasoning

Benchmark	Goal	Env	Comments	Paper
PHYRE	Make objects touch	2D	Games	Bakhtin et al., 2019
VirtualTools	Make objects touch	2D	Games	Allen, Smith, and Tenenbaum, 2020
OGRE	Make objects touch	2D	Games	Allen, Bakhtin, et al., 2020
CausalWorld	Build structures	3D	Robotic/RL	Ahmed et al., 2020
Phy-Q	Hit objects with slingshot	2D	RL (DQN baseline)	Xue et al., 2023
OPEn	Push objects (rel. positions)	3D	Continuous interaction	Gan et al., 2021
Language-Table	Push objects (abs. or rel. positions)	3D	Language input	Lynch et al., 2022
I-Phyre	Make objects touch	2D	Games	Li et al., 2024
SHM	Predict angle (pouring)	2D	Heuristic model	Li et al., 2025

Table 2: Puzzle Features: puzzle type - **Type**, if push action is available - **Push**, challenging scene elements (ramp - **Ramp**, narrow passages - **Walls**), and tool use divided into using available (**Use**) or creating new tools (**Build**). Pick/place and glue/unglue actions are available in all puzzles.

Id	Type	Push	Ramp	Walls	Tools	
					Use	Build
0,1,5,6	PnP	×	×	×	×	×
2-4	PnP	×	×	√	×	×
21,24,25	Glue	×	×	×	×	√
22,23,26	Glue	×	×	√	×	√
7	Push	√	×	√	√	×
8,12-15	Push	√	×	×	√	×
9,11	Push	√	×	×	√	√
10,27	Push	√	×	√	√	√
16,17,20	Dyn	√	√	×	×	×
18,19	Dyn	√	√	×	√	×

occurs: for instance, when the mouse is pressed to pick an object, or while moving the mouse. The frame data are saved every 10ms and contain the object poses, their parent objects, and connecting joints, such that the objects are represented as *frames* of the configuration. The data are used to analyze and cluster solutions, as well as to recreate the solution videos[1].

Solution Similarity and Strategy Clusters

Our approach for detecting strategies in solution attempts relied on trajectory similarity metrics, allowing us to cluster attempts and categorize solutions. In the course of our analysis, we evaluated different methods, including interaction sequences, state-based encodings, frame trajectories, and combinations of trajectories and action plans. Although action plans alone are sufficient for some puzzles, we found that they do not support the differentiation of solutions that vary in their trajectories.

To identify similar solutions for the same puzzle, we cluster complete solution trajectories using *Dynamic Time Warping* (DTW), which aligns time series of unequal length and is robust to differences in execution speed and temporal shifts.

[1]https://github.com/puzzles-cogsci/physical-puzzles

Dynamic Time Warping

Puzzle solutions (i.e., recorded frame data) are represented as positional vectors (i.e., the positions of all objects in a given puzzle). A solution with T time steps, K movable objects, and 2D coordinates is modeled as a multivariate time series $S = \{S_t^N\}, t \in \{1, \dots, T\}$,

$$S_t^N = (S_{1t}, S_{2t}, \dots, S_{Nt}), N = 2K \tag{1}$$

For two univariate solutions $S = \{S_{t_1}\}$ and $Q = \{Q_{t_2}\}$ with lengths T_1 and T_2, DTW computes their dissimilarity by finding an optimal warping path $W = \{(i, j)_k\}_{k=1}^K$ that minimizes the cumulative alignment cost:

$$DTW(S, Q) = \min_W \sum_{k=1}^K \delta(W_k), \tag{2}$$

where $\delta(W_k) = |S_i - Q_j|$ or $(S_i - Q_j)^2$ (Sakoe, 1978).

In our case, we consider the multidimensional input following the formulation in (Shokoohi-Yekta et al., 2017). For N-dimensional trajectories, we calculate the alignment cost using the *soft-DTW* distance metric (Cuturi and Blondel, 2017), which replaces the hard minimum with a differentiable soft-min operator with a smoothing parameter $\gamma \geq 0$:

$$DTW_\gamma = \min_W^\gamma \sum_{k=1}^K \delta(W_k), \tag{3}$$

with soft-min over $\{a_1, \dots, a_n\}$ defined as:

$$\min^\gamma\{a_1, \dots, a_n\} := \begin{cases} \min_i a_i, & \gamma = 0, \\ -\gamma \log \sum_{i=1}^n e^{-a_i/\gamma}, & \gamma > 0. \end{cases} \tag{4}$$

Identifying Strategy Clusters

We start by computing the DTW distances between the vectors of each attempt, forming a distance matrix D_i for a given puzzle i. The entries in the matrix D_i are the DTW alignment cost between two distinct attempts. Based on the constructed D_i, we perform hierarchical clustering with the Ward minimization algorithm (Ward, 1963). The identified clusters, their representative solutions and the variance of DTW scores, provide information about the applied strategies. For puzzles focusing on discrete decisions, cluster solutions differ in their action sequences. However, for puzzles with path restrictions, the positional differences may be more relevant than the order of interactions. Fig. 2 shows the interaction sequences for those two cases: while for Puzzle 1 the interactions are sufficient to distinguish the plans, for Puzzle 2, containing just one object, the paths taken are more important than the action plans.

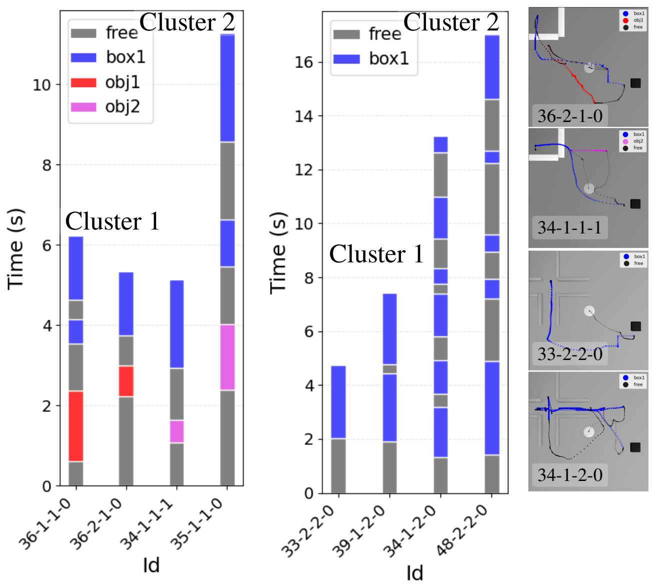


Figure 2: Example solutions in clusters and their trajectory heatmaps. In Puzzle 1 (left), solutions use different objects: `obj1` (red) in Cluster 1, and `obj2` (pink) in Cluster 2. In Puzzle 2 (center), solutions use different paths: upwards (direct) in Cluster 1, and from the right (with regrasps) in Cluster 2.

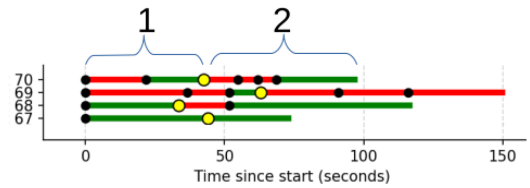
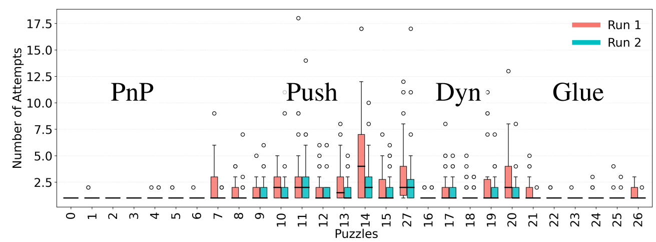
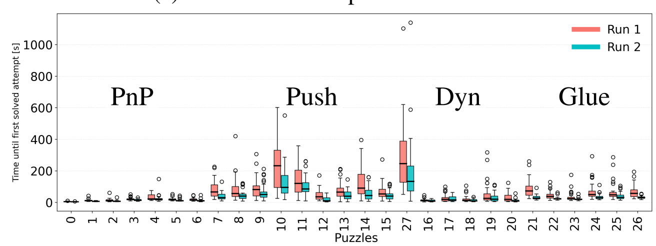


Figure 3: Successful (green) and failed (red) attempts for puzzle 15 for participants 67-70. *Time to solution* for participant 70: 1 for the 1st run, 2 for the 2nd run. Each black dot is a new attempt, the yellow dot is the start of the 2nd run.



(a) Number of attempts for both runs.



(b) Time-to-solve for both runs.

Figure 4: Attempts and time-to-solve metrics for both runs.

Results

We analyze the data to quantify the difficulty of the puzzles and compare participants' performance. We report the following metrics: (i) *time-to-solution* - the accumulated time of all attempts until the first solution during the same run (see Fig. 3), (ii) *average attempt time*, (iii) *difficulty* - the difficulty rating based on the survey, (iv) *diversity* - median DTW distance between attempts, (v) *repeatability* - average repeat of the same clusters, (vi) *success rate* - percentage of successful attempts, and (vii) *posterior failure rate* - expected number of failed attempts until the puzzle is solved. To compare the participants' success on a puzzle, we introduce the *performance* score as the ratio of the *time-to-solution* of a participant relative to the best *time-to-solution* for this puzzle.

Difficulty by Interaction Type

We analyze and report metrics of solution performance depending on puzzle type, to highlight the differences between interaction mechanics on solution difficulty. As seen in Fig. 4b, the puzzles requiring tool use and creation take more time to solve. The median time-to-solution for tool-use puzzles (7–15,18,19,27) is 65.4s vs 21.4s for the non-tool-use, with the tool-use&build subset (9-11,27) having the longest median time of 127.7s (IQRs [31.6-134.7], [11.7-43.5], and [72.2-268.3] respectively). Participants also tend to take more time to solve puzzles in the first run and are faster during the second: with the median time-to-solution decreasing from 35.5s to 21.1s (IQRs [16.2-74.8] and [10-42.8]). This suggests that once the participant finds a working solution, they may learn from that experience and are able to more efficiently implement the working strategy during the second run. In Fig. 4a, we report the number of attempts needed to solve a given puzzle. Here, we see that the Push and Dyn puzzles tend to require more attempts than PnP and Glue. In these puzzles, objects are more likely to be pushed into an inaccessible zone, forcing the participant to restart. Importantly, however, though Dyn puzzles require many attempts, the total time-to-solution is short compared to Push puzzles, which aligns well with the low difficulty rating in the surveys.

Reproducing Solutions

Another aspect we investigate is the difficulty to **repeat** a puzzle once it was **solved once**. We estimate the expected number of failures until a puzzle can be solved for the *first* time and the *second* time (i.e., typically during the second run

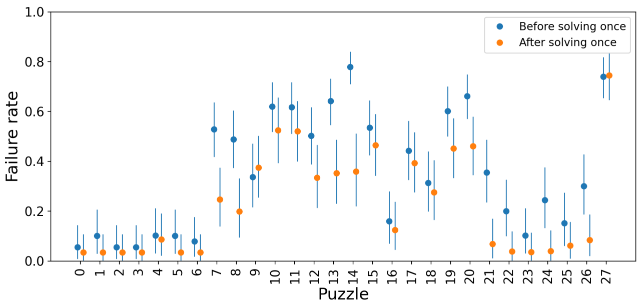


Figure 5: Failure rates before solving (blue) and after (orange).

of the puzzles). We assume that the failures follow a binomial distribution and compute the posterior mean failure rate using the overall failure rate for all puzzles as bias. Fig. 5 shows the rates before solving and after solving. Similarly to other metrics, PnP puzzles (0-6) are both easier to solve and easier to reproduce. The Glue puzzles (21-26), which do not allow pushing, are also easier to reproduce on average, even though they require more attempts than the PnP problems. The failure rates of Push puzzles have a less consistent pattern: while for some problems the rate improves significantly (7,8,12-14), for Tool-Build puzzles (9-11, 27) the improvement is much smaller or the failure rate actually slightly increases (9). Furthermore, the data from puzzles 13 and 14 show that some of the participants use gluing to move the target object, resulting in more reproducible strategies compared to solutions using pushing. In puzzle 15, the setup is almost the same as in puzzle 14 but the movable obstacle is replaced by a static one. We observe that though some participants still used gluing, some attempted to ricochet the object against the obstacle, a strategy which is more difficult to execute. For Dyn puzzles (16-20), the failure rate drop is smaller than for most Push puzzles, and for more difficult tasks (19,20) the failure rate improves *after solving* but remains around 50%.

Participant Performance

We found that the lower success rate of the attempts correlates with a lower performance score and higher repeats (Fig. 6). Participants with higher performance scores tended to have higher repeat scores. This is potentially because those participants were able to reproduce the strategy from the first solution. Participants with lower performance scores also had more repeating clusters but with lower success rates.

Puzzle metrics

For most puzzles, higher solution diversity is associated with longer attempts and higher difficulty ratings (Fig. 7b,7a). Exceptions are Push puzzles 13-15, which have a higher difficulty rating but a lower diversity score than Glue puzzles 24-26. Puzzles 13-15 are much more constrained and allow a smaller number of working solutions compared to puzzles 24-26, where the objects can be glued in many different ways to achieve the same result. Success rate also aligns with the difficulty rating: puzzles with lower success are on average

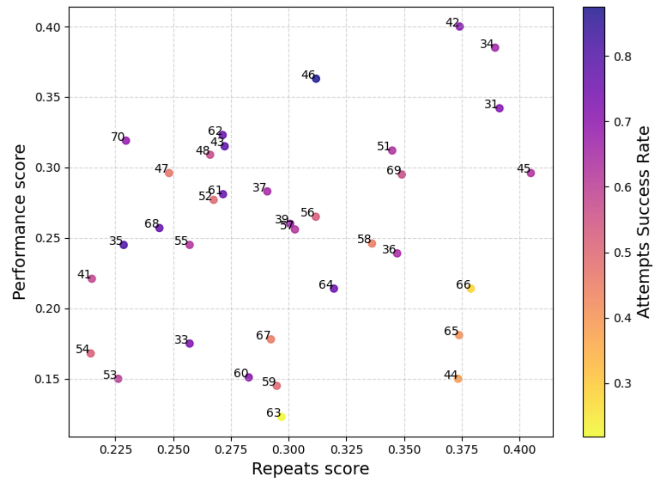


Figure 6: Performance and cluster repeat scores of participants who completed both runs (n = 35).

rated as harder (Fig. 7c). For some Dyn puzzles (17, 19 and 20), the participants required many attempts but did not perceive them as difficult. For those puzzles, the attempts are much shorter and the time-to-solution is closer to the easier puzzles, despite the higher failure rate. Overall, the time-to-solution (shown as a color gradient in Fig. 7) is strongly correlated with the difficulty rating (r = .938, p < .001).

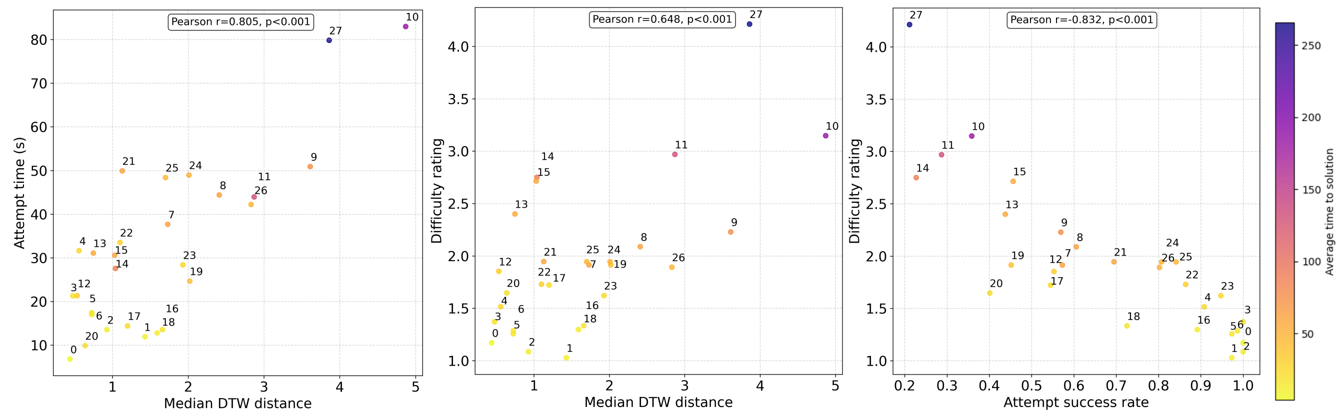
Strategy Clusters by Puzzle Type

In addition to the overall diversity of solutions applied to each puzzle, we analyze the diversity of solution attempts inside the identified clusters. We report our findings in Fig. 8 for three puzzle groups: *Pushing with Tool-Use*, *Dynamic* (18-19 with Tool-Use), and *Gluing* with narrow walls. For most Tool-Use puzzles (12,14,18,19), we see that successful solutions tend to build tighter clusters (with similar solutions), and failed attempts are more spread out. This indicates that unsuccessful attempts in these problems have a greater variation in trajectories, either due to additional movements or due to failed interactions. Dynamic puzzles 16,17 and 20 have low diversity, and the clusters are clearly separated into working and failed solutions: even if they require several attempts, the attempts are repeated with only small changes. Finally, the Glue puzzles allow for many different working strategies, and their solution attempts form several clusters with a high success rate but high DTW scores.

Conclusion

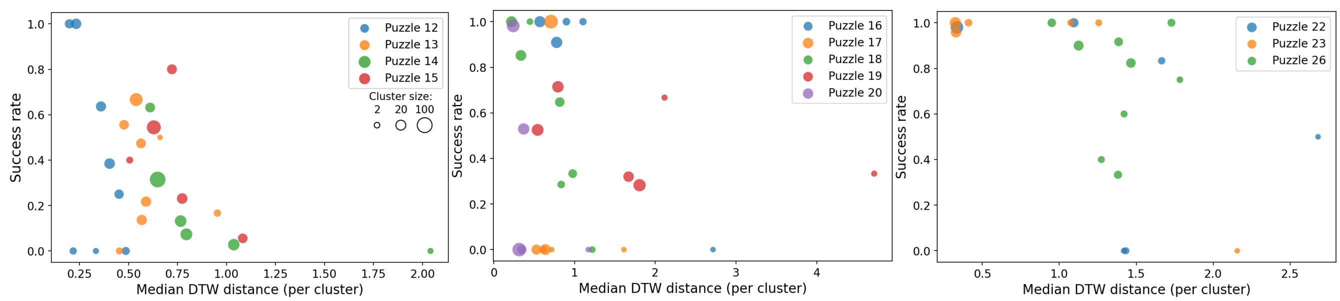
We introduced a set of 28 virtual physics puzzles requiring long-horizon sequential manipulation with pick-and-place, pushing, and tool use/creation, and used collected data to quantify performance and identify strategies, addressing the following research questions: (i) puzzle difficulty, (ii) strategy discovery and switching, and (iii) impact of exploration and repetitions on solution performance.

(i) We found that puzzle difficulty varies systematically with



(a) Diversity vs attempt time. (b) Diversity vs difficulty rating. (c) Attempt success rate vs difficulty rating.

Figure 7: Diversity, attempt length, success rate and difficulty rating for all puzzles. The color shows the average time-to-solution.



(a) Push puzzles with tool-use (b) Dyn puzzles (c) Glue puzzles with walls.

Figure 8: Median DTW distances and success rates of clusters, excluding clusters of size=1. Higher success rates in clusters correlate with higher attempt similarity (low DTW distances) for Push-Tool-Use (left) and Dyn-Tool-Use Puzzles (18-19, center). Solutions in Dyn Puzzles without Tool-Use (16, 17, 20) only require parametric changes and have an overall high similarity. Solutions for Glue Puzzles (right) have many distinct clusters, but with an overall high success rate.

task features: puzzles that combine tool creation with pushing are the most challenging, with longer time-to-solution, more attempts, and higher failure rates, consistent with a larger combinatorial search space and a more challenging control. The hardest puzzles (10 and 27), in addition to tool creation and pushing, included geometric restrictions, such as narrow passages. This suggests that the most difficult puzzles are those that allow for higher combinatorial complexity in action selection and require precise positioning and control. Overall, puzzles with controlled manipulation such as PnP and Glue were easier to solve and easier to reproduce after solving once, compared to puzzles involving pushing.

(ii) Using trajectory-based clustering, we were able to identify tight clusters of similar solutions within most of the puzzles, suggesting that participants converge to similar plans. The more challenging puzzles with multi-step object interactions (especially Tool-Use) show a higher diversity of solutions, with strategies making use of diverse mechanics. For Dynamic puzzles that require fewer steps, the solutions have a higher similarity and can be refined by small parametric changes.

(iii) For all puzzles, we observed an improvement **after** the participant had solved the puzzle once (in time and number of attempts). The ability to repeat a **working** solution leads to a higher performance score: among participants with similar success rate, those with the high repeat score achieved a better performance, indicating that their repeated solutions were likely faster. Overall, we found that some problems are more difficult to reproduce than others: while PnP/Glue puzzles have a high success rate for all participants, repeating solutions for Push/Dyn puzzles was more likely to fail. In particular, the outcome of pushing an object is harder to predict, and applying a pushing strategy with less controlled movements may result in more variation of trajectories and therefore higher diversity scores. In cases where the solution did not work, but the puzzle could still be solved, many participants preferred to continue the attempt, rather than restart it and repeat the previous steps. This demonstrates the flexibility of human problem solving, as participants are able to continue their solution from a completely new state instead of following the initial plan. However, this also indicates that repeating actions is associated with a cost and the preference for recovering (rather than resetting) might depend on the control demands of the interaction mechanics and the uncertainty about the potential success for a given participant or problem.

Acknowledgments

This research has been supported by the German Research Foundation (DFG) under Germany's Excellence Strategy – EXC2002/1–390523135 "Science of Intelligence".

References

Ahmed, O., Träuble, F., Goyal, A., Neitz, A., Bengio, Y., Schölkopf, B., Wüthrich, M., & Bauer, S. (2020). Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. https://doi.org/10.48550/arXiv.2010.04296

Allen, K. R., Bakhtin, A., Smith, K., Tenenbaum, J. B., & van der Maaten, L. (2020). Ogre: An object-based generalization for reasoning environment. *Neurips Workshop Object Representation for Learning and Reasoning*.

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), 29302–29310. https://doi.org/10.1073/pnas.1912341117

Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L., & Girshick, R. (2019). Phyre: A new benchmark for physical reasoning. *Advances in Neural Information Processing Systems*, *32*, 5082–5093.

Coumans, E. (2015). Bullet physics simulation. *ACM SIGGRAPH 2015 Courses*. https://doi.org/10.1145/2776880.2792704

Cuturi, M., & Blondel, M. (2017). Soft-dtw: A differentiable loss function for time-series. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 894–903.

Dörner, D., & Funke, J. (2017). Complex problem solving: What it is and what it is not. *Frontiers in Psychology*, *8*, 1153.

Gan, C., Bhandwaldar, A., Torralba, A., Tenenbaum, J. B., & Isola, P. (2021). Open: An open-ended physics environment for learning without a task. https://doi.org/10.48550/arXiv.2110.06912

Karimpur, H., Wolf, C., & Fiehler, K. (2024). The (un)ideal physicist: How humans rely on object interaction for friction estimates [PMID: 38252798]. *Psychological Science*, *35*(2), 191–201. https://doi.org/10.1177/09567976231221789

Köhler, W. (1925). *The mentality of apes*. Harcourt, Brace & Company.

Li, S., Ma, Y., Yan, J., Dai, B., Peng, Y., Zhang, C., & Zhu, Y. (2025). A simulation-heuristics dual-process model for intuitive physics. *CogSci*.

Li, S., Wu, K., Zhang, C., & Zhu, Y. (2024). I-PHYRE: interactive physical reasoning. *The Twelfth International Conference on Learning Representations, ICLR 2024*.

Lozano-Pérez, T., & Kaelbling, L. P. (2014). A constraint-based method for solving sequential manipulation planning problems. *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3684–3691.

Lynch, C., Wahid, A., Tompson, J., Ding, T., Betker, J., Baruch, R., Armstrong, T., & Florence, P. (2022). Interactive language: Talking to robots in real time. https://doi.org/10.48550/arXiv.2210.06407

Mitko, A., & Fischer, J. (2020). When it all falls down: The relationship between intuitive physics and spatial cognition. *Cognitive research: principles and implications*, *5*, 1–13.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.

Sakoe, H. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*, 159–165. https://api.semanticscholar.org/CorpusID:17900407

Shokoohi-Yekta, M., Hu, B., Jin, H., Wang, J., & Keogh, E. (2017). Generalizing dtw to the multi-dimensional case requires an adaptive approach. *Data mining and knowledge discovery*.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review Monograph Supplement 2*, *4*(Whole No. 8).

Toussaint, M. A., Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2018). Differentiable physics and stable modes for tool-use and manipulation planning. *Proceedings of the Robotics: Science and Systems*.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Xue, C., Pinto, V., Gamage, C., Nikonova, E., Zhang, P., & Renz, J. (2023). Phy-q as a measure for physical reasoning intelligence. *Nature Machine Intelligence*, *5*(1), 83–93.

Yildirim, I., Gerstenberg, T., Saeed, B., Toussaint, M., & Tenenbaum, J. (2017). Physical problem solving: Joint planning with symbolic, geometric, and dynamic constraints. *arXiv preprint arXiv:1707.08212*.