Knot So Simple:
A Minimalistic Environment for Spatial Reasoning

Zizhao Chen & Yoav Artzi

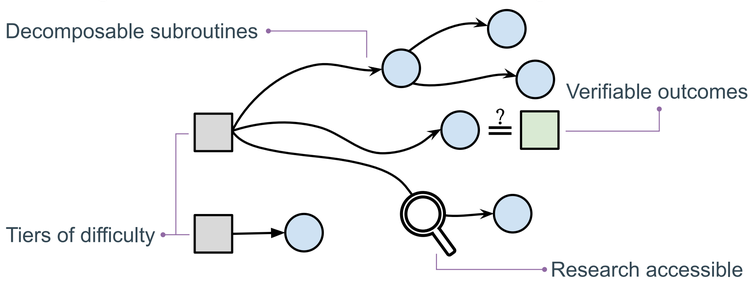
🤔 Can you convert  to  without breaking the rope? How?

Natural language is **not** all you need

Spatial concepts like pose and geometry are hard to reason about in natural language alone. Models need visual reasoning.
But, existing reasoning benchmarks focus on text-only reasoning.

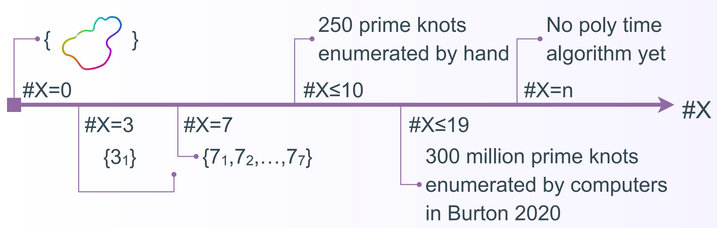
Desiderata: Shapes of reasoning

What makes MATH/AIME successful benchmarks for text-only reasoning? How does it transfer to visual reasoning?



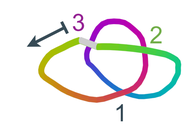
💡 **Knots** are **constrained** in their perception space yet **expressive** in their reasoning structure → Perfect for visual reasoning

KnotGym is a special case of the (open!) **knot equivalence problem**. There are many possible Gauss Codes at given #X=n. Are they all continuously convertible between each other? No!

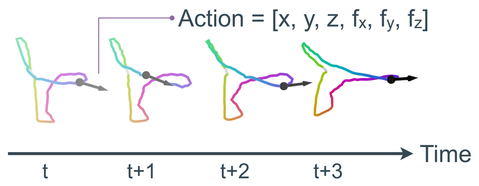


Design 1: KnotGym is about rope manipulation

An agent is shown **an initial knot** image and **an goal knot** image, and outputs a series of **actions** to convert the knot to the goal, as measured by their **Gauss Codes** (GCs), i.e., GC(current) = GC(goal).



Gauss Code:
+1, -2, -3, -1, +2, +3



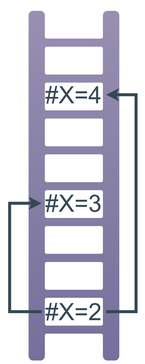
Design 2: Three task families

Task	Description	Example task observations (Initial, Goal) → (Final, Goal)
`unknot`	Untangle a knot into a simple loop (#X=n→0)	
`tie`	Tie a goal knot from a simple loop (#X=0→n)	
`convert`	Tie a new knot from an old knot (#X=n_i→n_g)	

Design 3: A difficulty ladder by #X for generalization

Despite continuous state space, formalizations help:
- **Verifiable outcome** via Gauss Code
- **Quantifiable complexity** via the crossing number (#X)

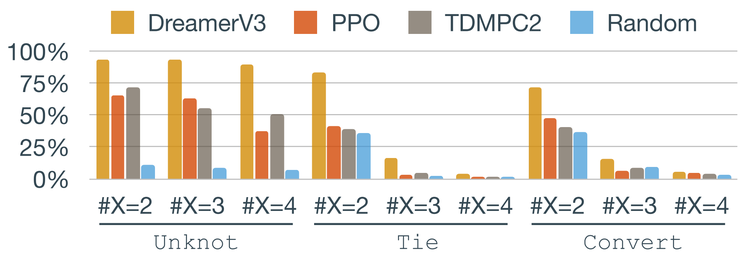
They enable:
- **A ladder of task difficulty**, e.g., `convert` #X=2, `tie` #X=3
  - Brute-force proof: factorial goal space GCs wrt. #X
- Defining **generalization** by #X
  - Train vs. test split at #X=2
  - Train on #X=2 and test on #X=4



Design 4: Familiar and accessible with MuJoCo

- Gym interface; VecEnv on CPUs
- Implemented with MuJoCo; extensible like adding a camera etc.

Experiment: RL

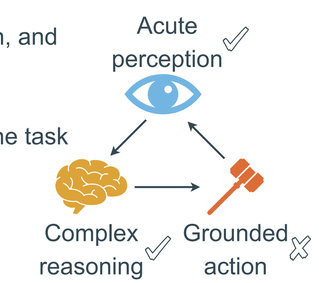


Train success rate of comparable sized policies over 1M steps

- Task difficulty reflected by #X: 2 < 3 << 4
  - Expected as the goal and state space grows wrt. #X
- `tie` and `convert` are much harder than `unknot`
  - For `unknot` a trivial solution exists
  - **Goal-conditioning is hard** because of large goal space
- Preliminary analysis on train-test and cross #X generalizations
  - Lacking effective exploration

Experiment: Prompting VLMs

We prompt GPT4o about the task domain, and apply the generated action tuples
- Below random average success rate
- Correctly perceives + reasons about the task
- **But ungrounded actions**
  - Even when history is included
  - Elaborative visual guidelines help



Future work

- Multi-modal reasoning by generating image/latent reasoning tokens
- Planning with learned neural/programmatic world models
- Learning physics and subroutines through interaction

More at lil-lab.github.io/knotgym

- Rollout examples
- Error analysis
- Training curves