PhysIQ

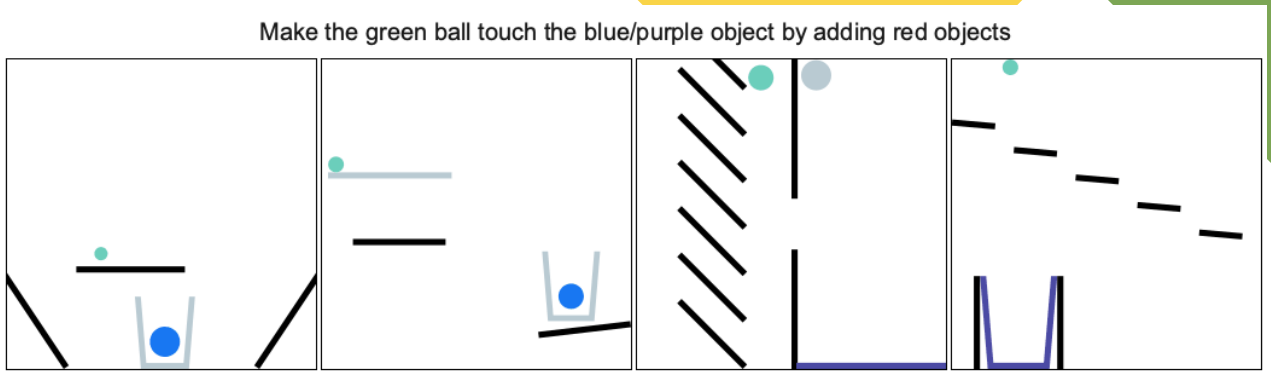
A Physical Reasoning Bechnmark for VLMs

By Massimo Stefan

Index

Introduction

Physical Reasoning



"The ability to predict and understand the interactions, dynamics, and outcomes within physical environments, based on causal relationships and the laws of physics"

🔮 **Prediction and Understanding:**

- Anticipating interactions, dynamics, and outcomes based on causal relationships and physical laws.

🧠 **Foundational Aspects:**

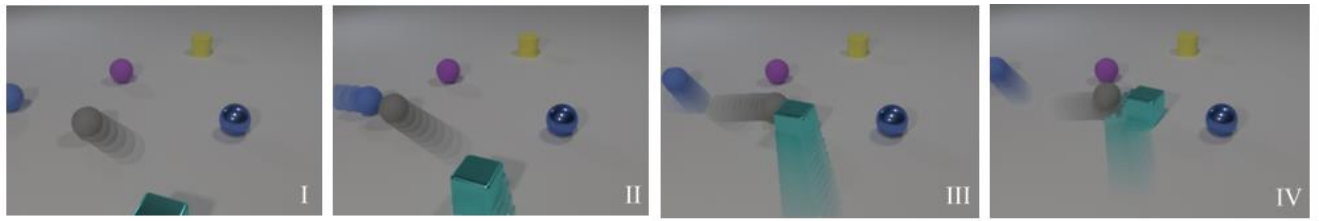
- Causal Reasoning: Differentiating between cause-and-effect relationships and superficial correlations.
- Mental Simulations: Humans intuitively simulate scenarios internally to anticipate outcomes.

❓ **Why It Matters in AI:**

- Essential for robotics, virtual assistants, and real-world interactive systems.
- Ensures reliable predictions in real-world interactions beyond learned patterns.

Several benchmarks

SHAPES



Descriptive:

Q: *How many spheres are moving?*
A: *2*

Q: *What shape is the second object to collide with the gray object?*
A: *Cube*

Q: *Are there any collisions after the cube enters the scene?*
A: *Yes*

Explanatory:

Q: *Which of the following is responsible for the collision between the gray object and the cube?*

a) The presence of the purple object
b) The collision between the blue sphere and the gray sphere
c) The presence of the purple object
d) The presence of the blue object

A: *b), d)*

Predictive:

Q: *What will happen next?*

a) The cube and the gray object collide
b) The gray sphere collides with the purple sphere
c) The metal sphere and the cube collide
d) The gray sphere collides with the blue sphere

A: *b)*

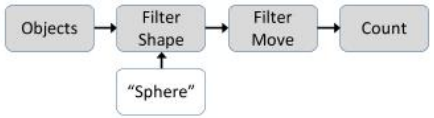
Counterfactual:

Q: *What will happen if the gray sphere is removed?*

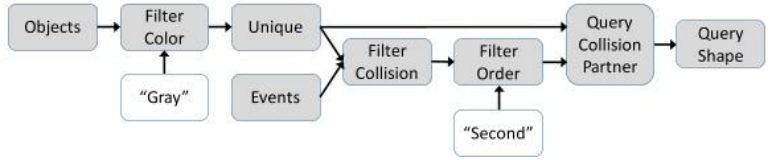
a) The blue sphere collides with the cube
b) The blue sphere and the metal sphere collide
c) The purple object collides with the cylinder
d) The cube and the metal sphere collide

A: *a), d)*

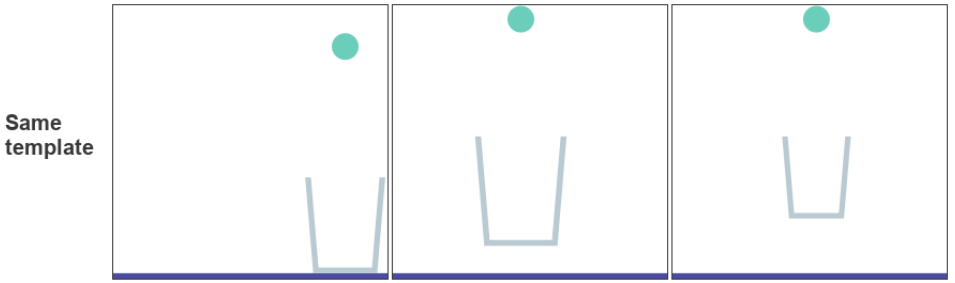
How many spheres are moving?



What shape is the second object to collide with the gray object?

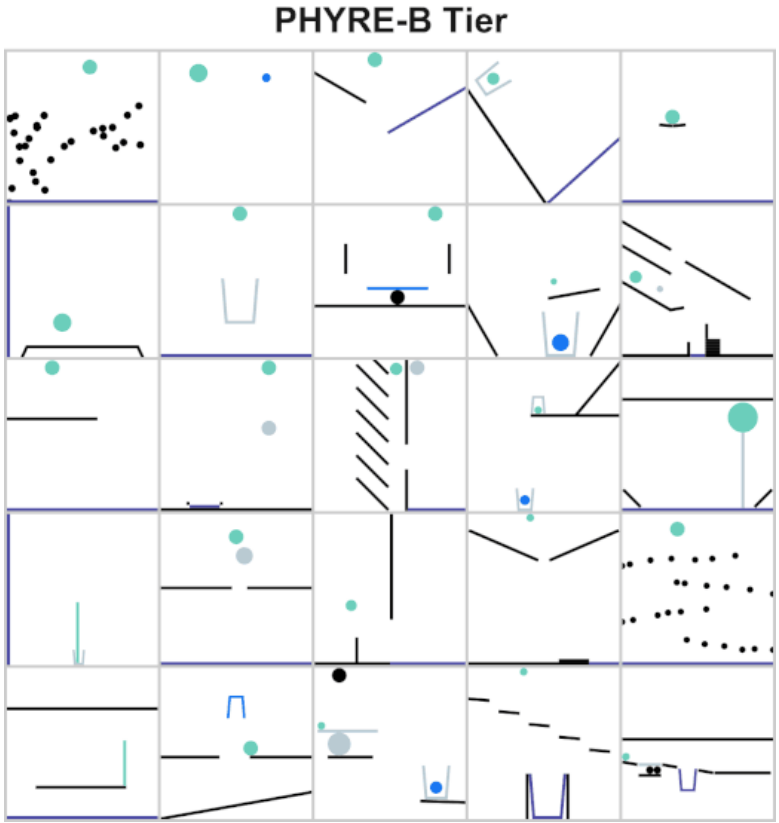


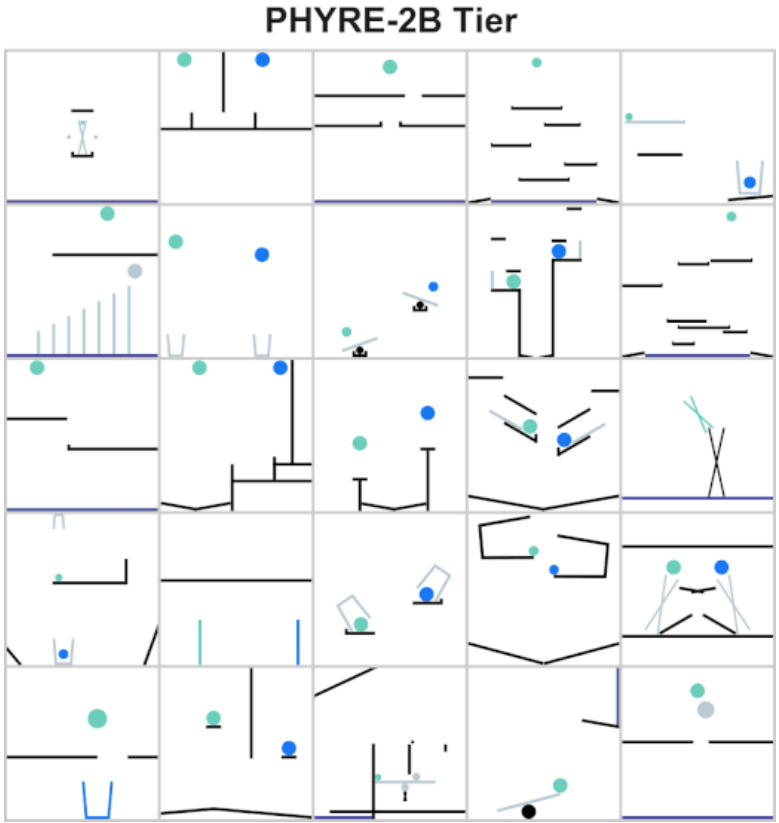
The PhyRE Dataset



# of samples:

- 25 “1 ball” templates
- 25 “2 balls” templates
- 100 Iterations per template
- Total: 5000 different puzzles





PhysIQ: a PhyRE extension

📏 Measure the following abilities in the physical domain:

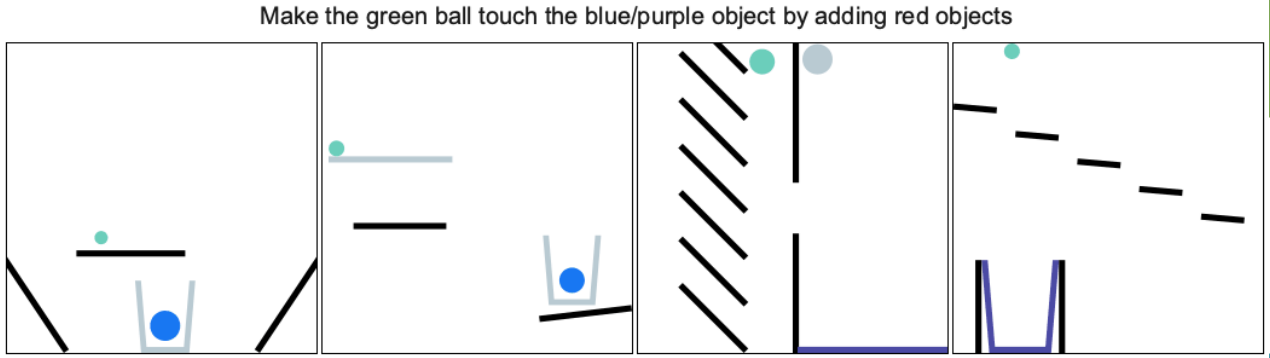
- Understanding
- Confidence
- Prediction
- Comparison
- Interactivity

✅ Pros:

- Extensible
- Customizable
- Offline & Interactive
- 2D
- Simple for an LLM to act on

Methodology

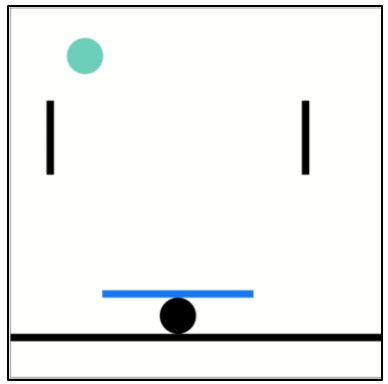
Puzzles Bodies

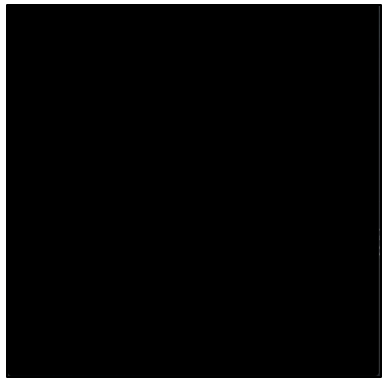


1. **Position**: The (x, y) coordinates of the body's center within the scene.
2. **Body Type**: Determines the physics behavior of the object:
    1. **Static** (0): Fixed objects that do not move during simulation.
    2. **Dynamic** (1): Objects that respond to forces and collisions.
3. **Angle**: The rotation of the body in radians.
4. **Relationship**: Define the goal condition
5. **Color:** An integer index that maps to a specific color:
    1. **Red** (0)
    2. **Black** (1)
    3. **Green** (2)
    4. **Azure/Blue** (3)
    5. **Purple** (4)
    6. **Grey** (5)
6. **Shape Type**: Defines the geometric form of the body:
    1. **Polygon** (0, 2): Defined by a set of vertices.
    2. **Circle** (1): Defined by a center point and radius.
    3. **Compound** (3, 4): Complex shapes composed of multiple polygons.

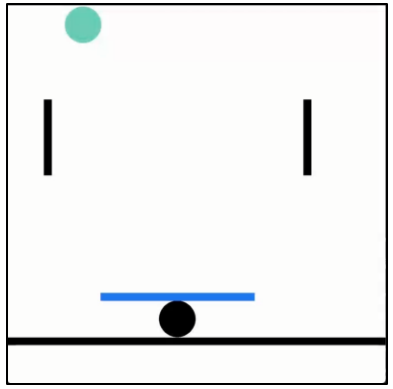
Simulation Constants

Original





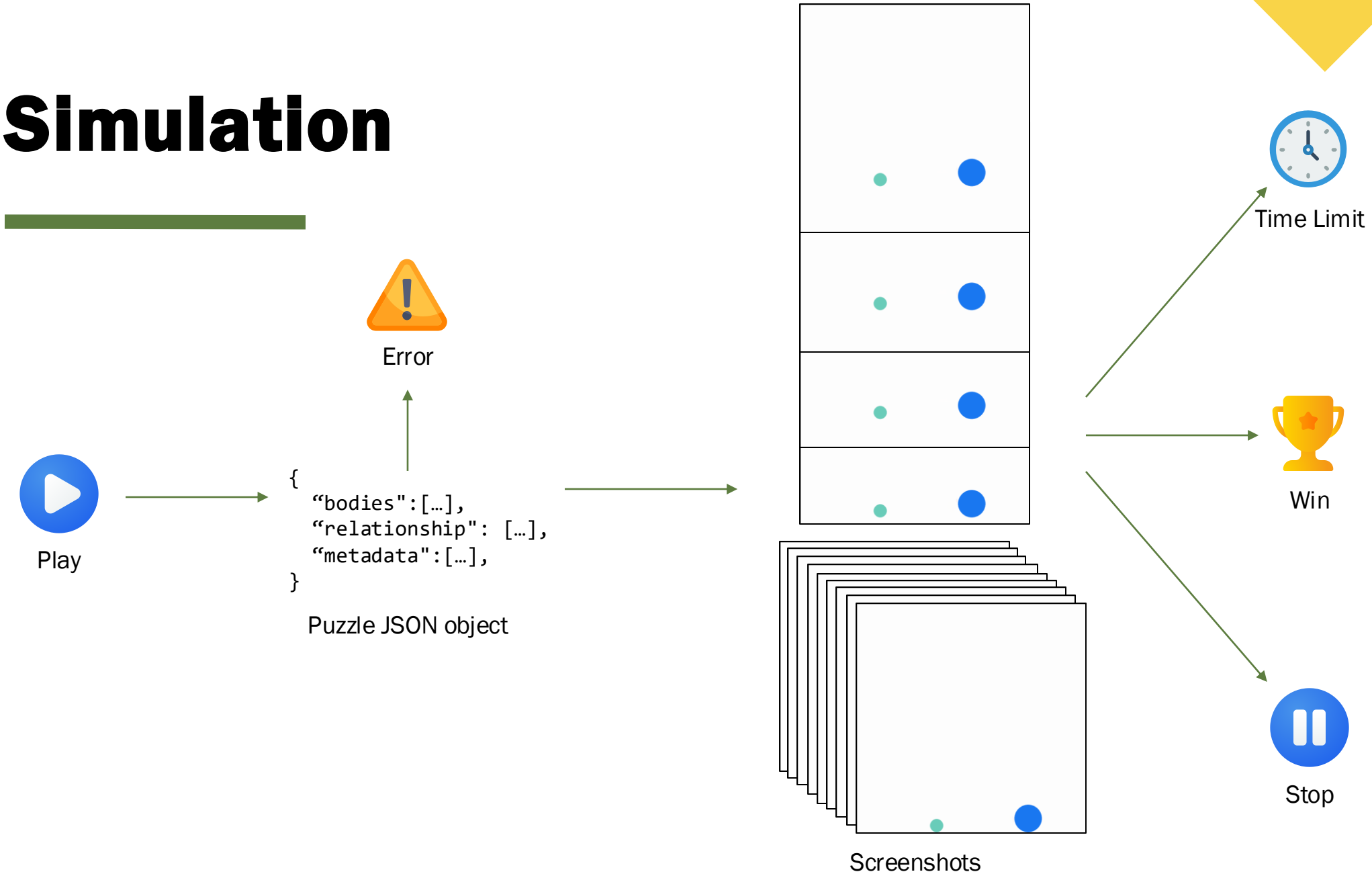
Pymunk



PyBox2D

Category	Parameter	Description
Physical	Scene dimensions	Size of the simulation scene (256 × 256 pixels)
	Gravity	Acceleration due to gravity (9.81 m/s^2)
	Density	Mass per unit area (0.25 kg/m^2)
	Friction coefficient	Coefficient determining resistance to sliding (0.5)
	Elasticity (restitution)	Bounciness of collisions (0.20)
	Angular damping	Reduction factor for rotational motion (0.01)
	Linear damping	Reduction factor for linear motion (0.0)
	Min Proposal Radius	Minimal length of the proposals (2)
	Max Proposal Radius	Maximum length of the proposals (32)
Simulation	Frame rate	Number of frames per second (60 FPS)
	Time scale	Factor to adjust simulation speed (1.0)
	Scene dimensions	Rendered scene dimensions (256 × 256 pixels)
	Velocity iterations	Iterations for the velocity solver (10)
	Position iterations	Iterations for the position solver (10)
Stopping	Stop velocity threshold	Threshold below which an object is considered static (0.1)
	Required frames for early stop	Consecutive frames with static objects to trigger early stopping (400)
	Required frames for goal verification	Consecutive frames where the two target objects had to remain in contact (360)
	Max frames	Maximum frames allowed for simulation before forced termination (3000)

Simulation


Error
Play
{
“bodies":[…],
“relationship": […],
“metadata":[…],
}
Puzzle JSON object
Screenshots
Time Limit
Win
Stop

Proposals Identification

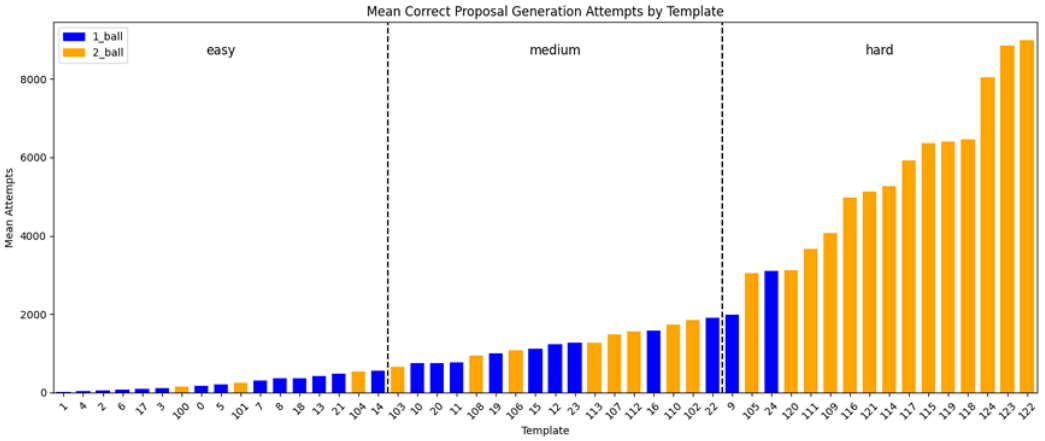


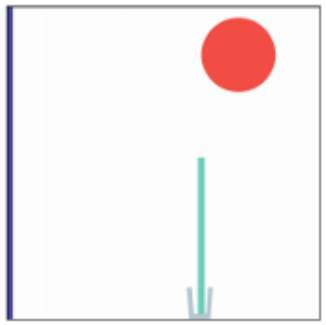
Figure 4.3: Mean Correct Proposal Generation Attempts by Template.

🎯 **Goal**: create positive and negative samples to test the VLMs

⚙️ **Properties**:

- 10k attempts per puzzle
- 20 correct proposals per template
- 3 incorrect proposal per correct proposal:
  - 1 hard: displacement from 0 to 1 original radius
  - 1 medium: displacement from 1 to 2 original radius
  - 1 easy: displacement from 2 to 4 original radius

  + new radius between $\frac{1}{2}$ and $\frac{3}{2}$ of the original radius



(a) Correct

Figure 4.4: The 4 proposals found for template 00003 iteration 000.

Experiments & Results

VLMs Evaluated

Name	Provider	Size	Weights
Pixtral-Large-2411	MistralAI	123B	Open
Qwen2.5-VL-72B-Instruct	Qwen	72B	Open
Llama-3.2-90B-Vision-Instruct	Meta	90B	Open
Llama-3.2-11B-Vision-Instruct	Meta	11B	Open
Grok-2-Vision-1212	xAI	Big	Closed
Gemini-2.0-Flash	Google	Small	Closed
Claude 3.5 Sonnet	Anthropic	Big	Closed
GPT-4o	OpenAI	Big	Closed

💰 To reduce the evaluation costs, we selected only the best performing VLMs for the harder challenges:

- Sanity Check (8 VLMs)
- Static Evaluations (4 VLMs)
- Interactive Evaluation (2 VLMs)

Prompts design

System prompt

User prompt

Assistant Answer

The prompts are always composed by:

- A **System message**
  - **Role** (e.g. "…physics expert analyzing…)
  - **Simulation parameters** (optional)
  - **Task description**
  - **Response format**
- **Few-shot examples** (optional) each composed by:
  - **A User message** with the example (randomly sampled)
  - **An Assistant message** with the correct response
- **A User message**:
  - With the actual question

…
Simulation conditions for all tasks:
- Gravity: 9.81m/s²(downward)
- Objects density: 0.25
- Friction coefficient: 0.5
- Elasticity coefficient: 0.2
- Simulationduration:25 seconds or until objects stop moving
- Black and purple objects: static (fixed)
- Green, red, grey objects: dynamic (can move)
- Goal criterion: Objects must remain in contact ≥ 3 seconds
- Objects cannot leave the visible simulation boundaries

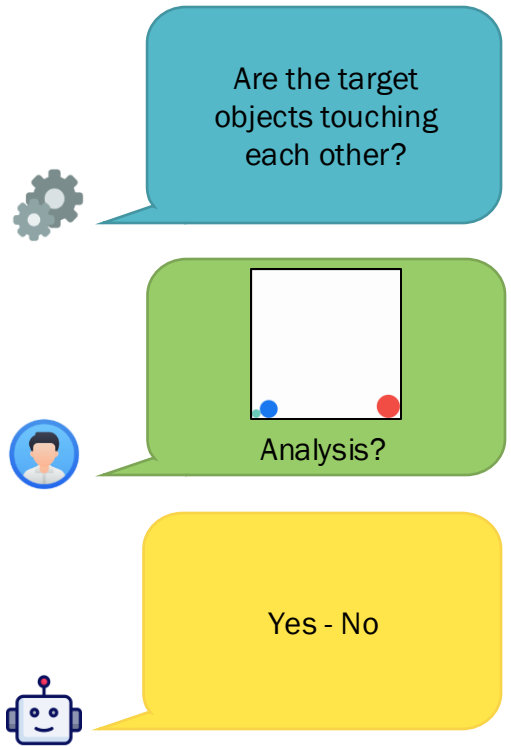
…

…
Clearly define your solution by specifying:
- "x": horizontal position of the ball center (0 is left, maximum is 256 on the right)
- "y": vertical position of the ball center (0 is bottom, maximum is 256 at the top)
- "radius": size of the ball (minimum 2,maximum 32)

Important rules for placing the ball:
- The ball must remain fully within the visible simulation boundaries.
- The ball cannot overlap with existing objects.

…

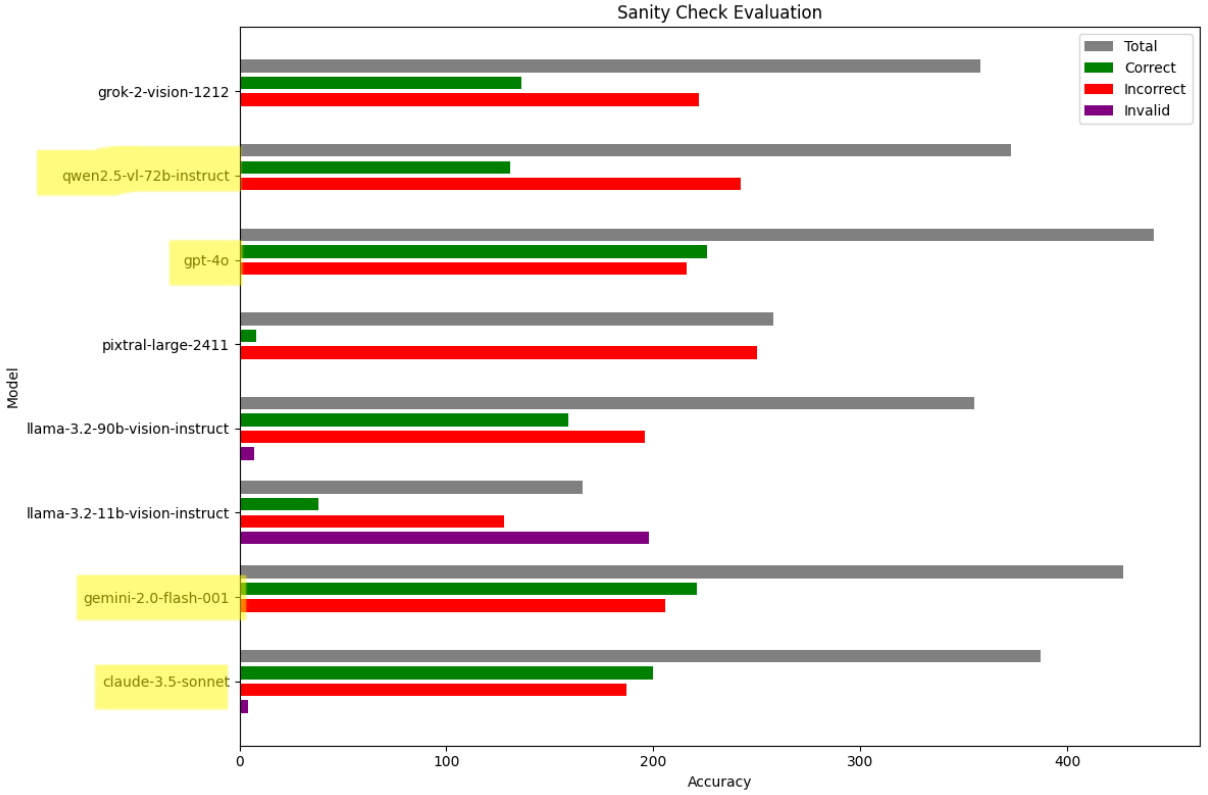
Sanity Check Evaluation



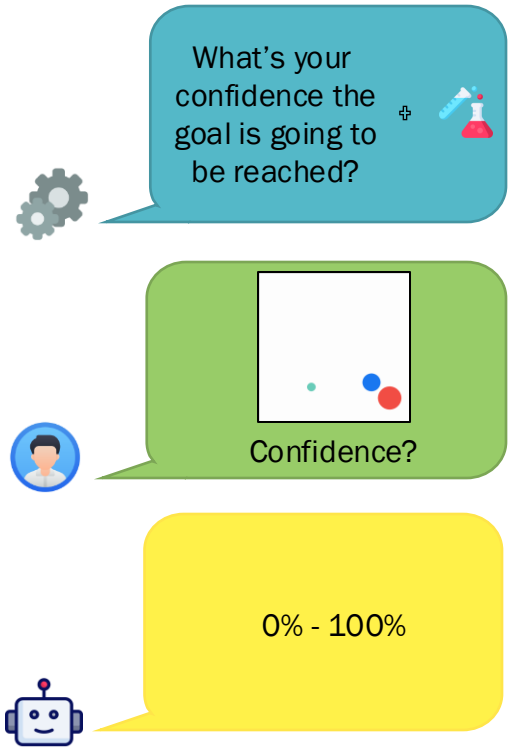
Properties:

- # samples per model: 5 correct & 5 incorrect proposals per template (500 total)
- No few-shot examples
- Only the 4 best performing VLMs proceed in the evaluation

Goal: determine weather the model can understand if the target objects are in contact or not



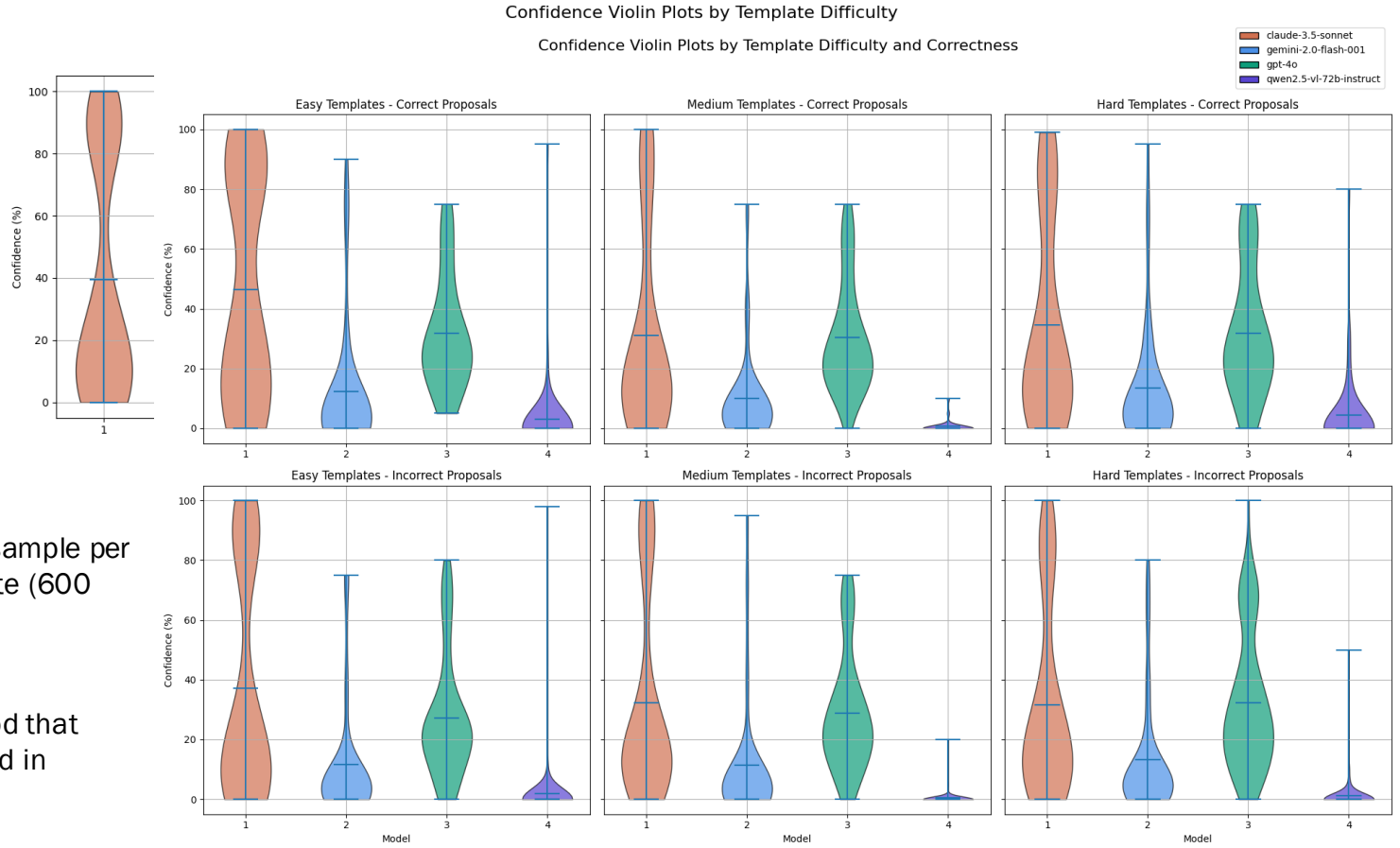
Confidence Estimation



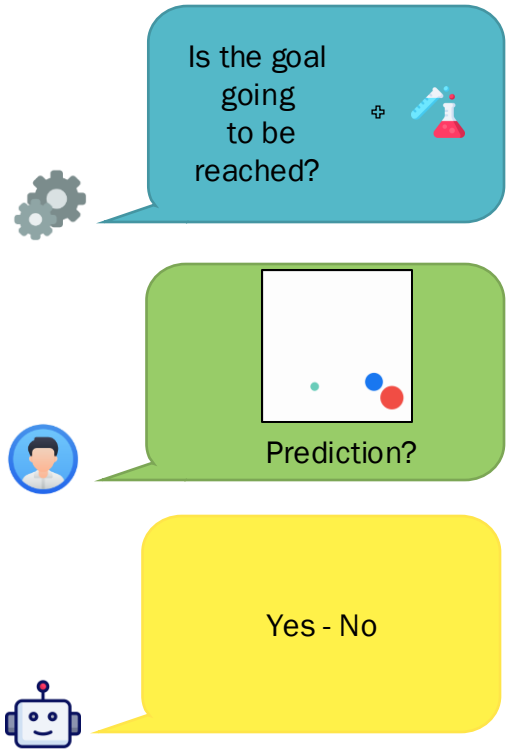
Properties:

- # samples per model: 3 sample per proposal type per template (600 total)
- No few-shot examples

Goal: attest the likelihood that a given proposal will succeed in the goal accomplishment



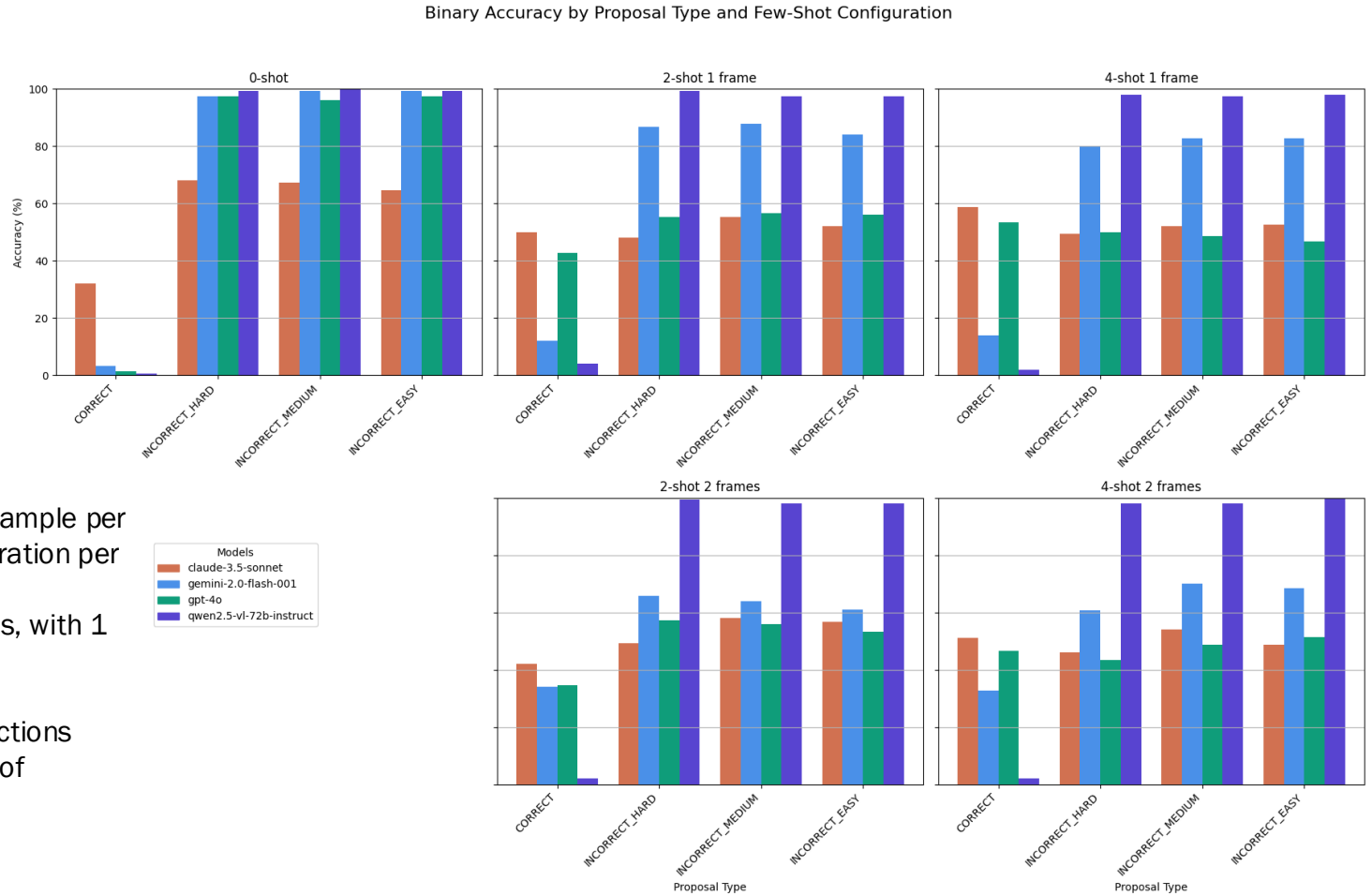
Binary Classification



Properties:

- # samples per model: 3 sample per proposal type per configuration per template (3000 total)
- 0-, 2- and 4-shot examples, with 1 or 2 frames each

Goal: measure the predictions abilities with different types of contexts



Ranking Task

What's the proposal's ranks from most to least likely to accomplish the goal?

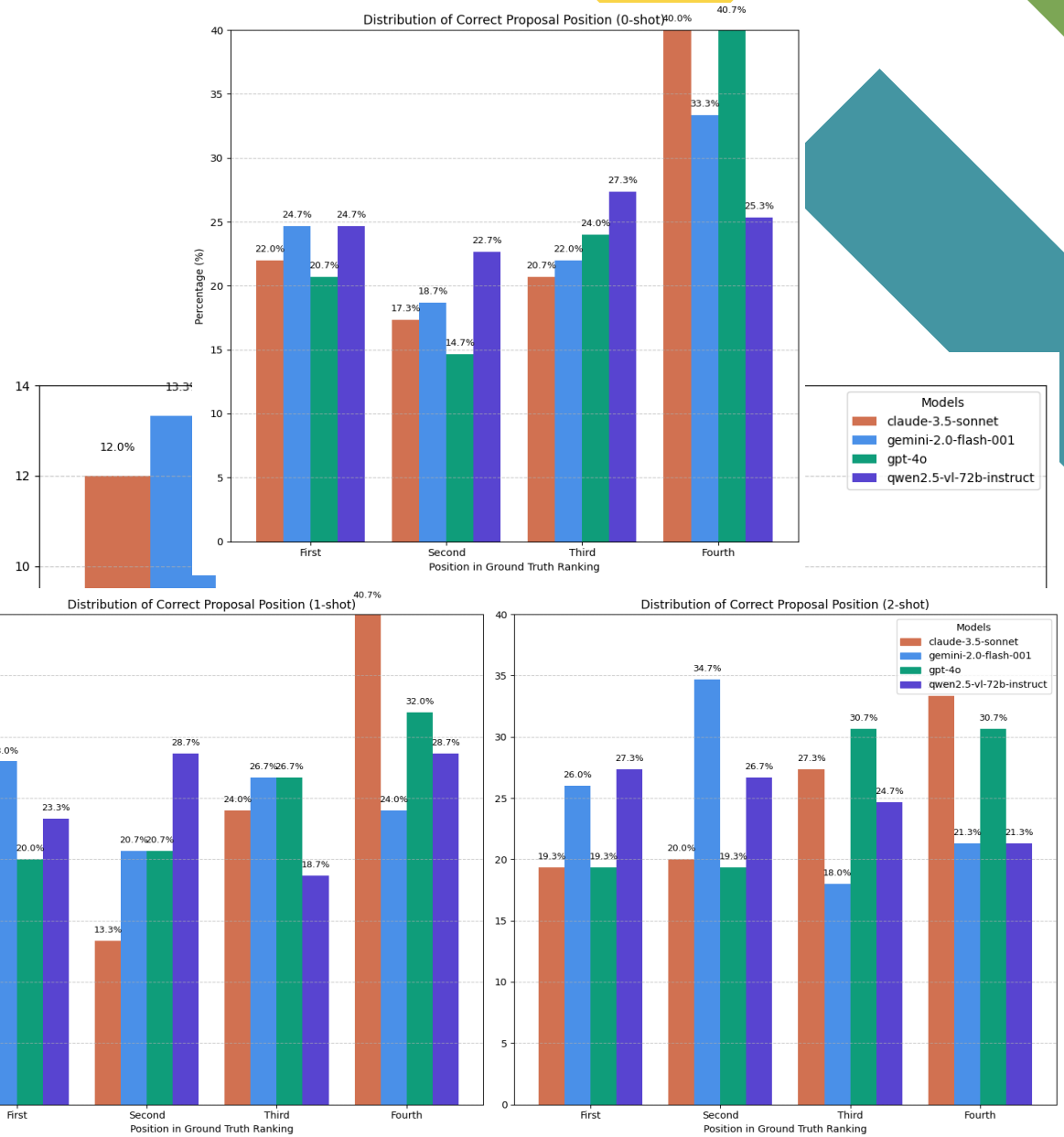
Comparison?

[1, 3, 4, 2]

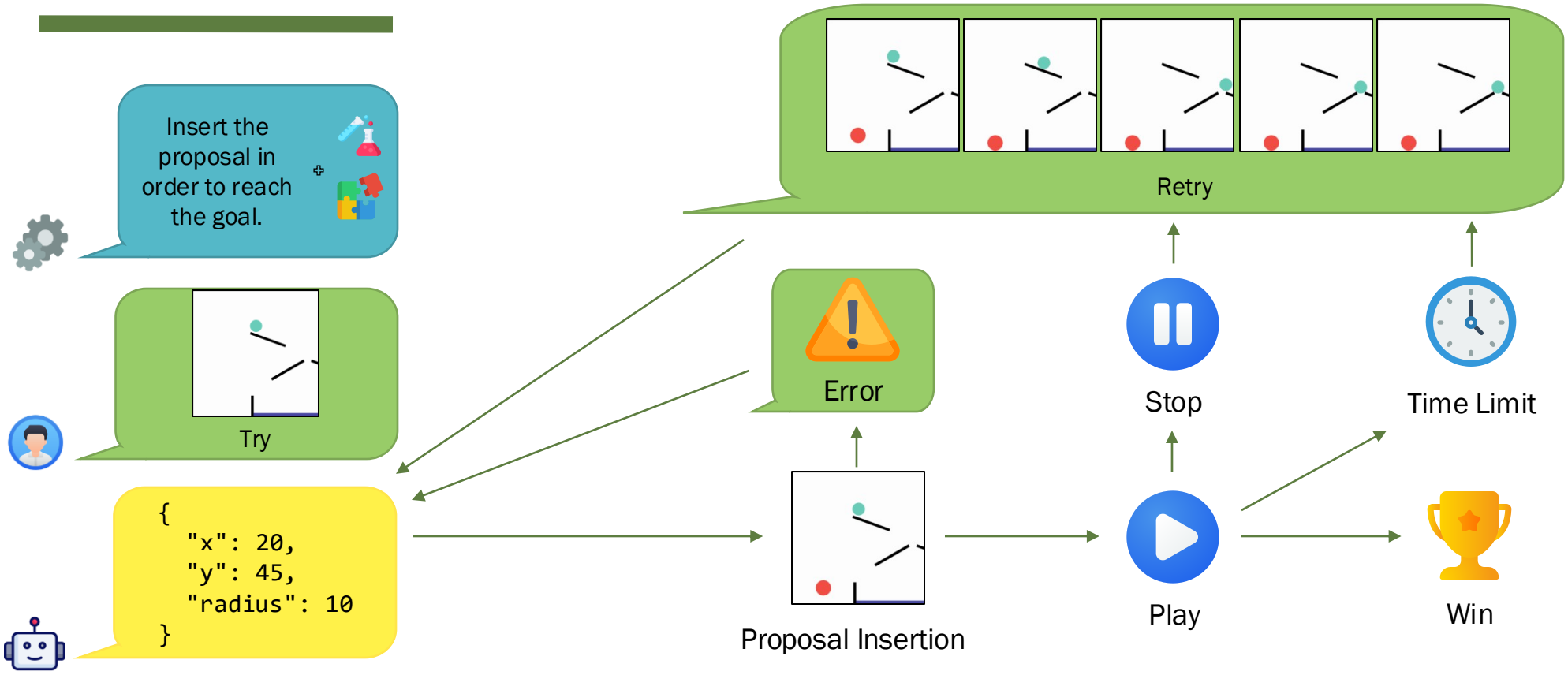
Properties:

- # samples: 3 samples per template per configuration (450 total)
- 0-, 1- and 2-shot examples

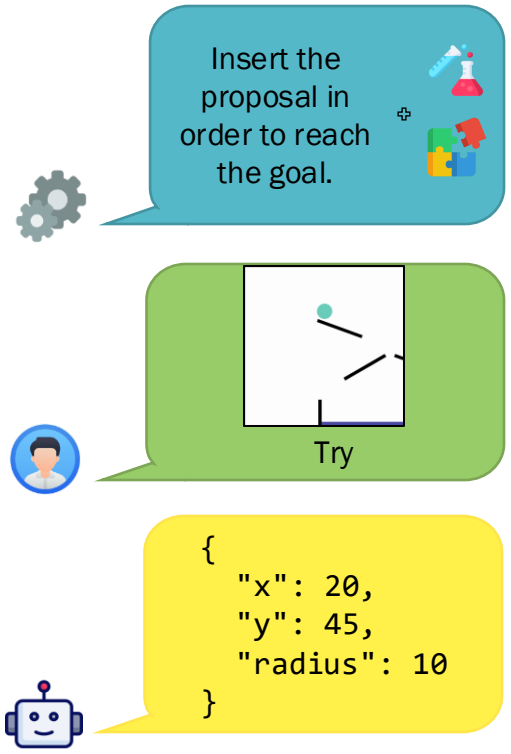
Goal: determine weather the model can rank the proposals by their likelihood of successfully meeting the puzzle's goal



Interactive Insertion



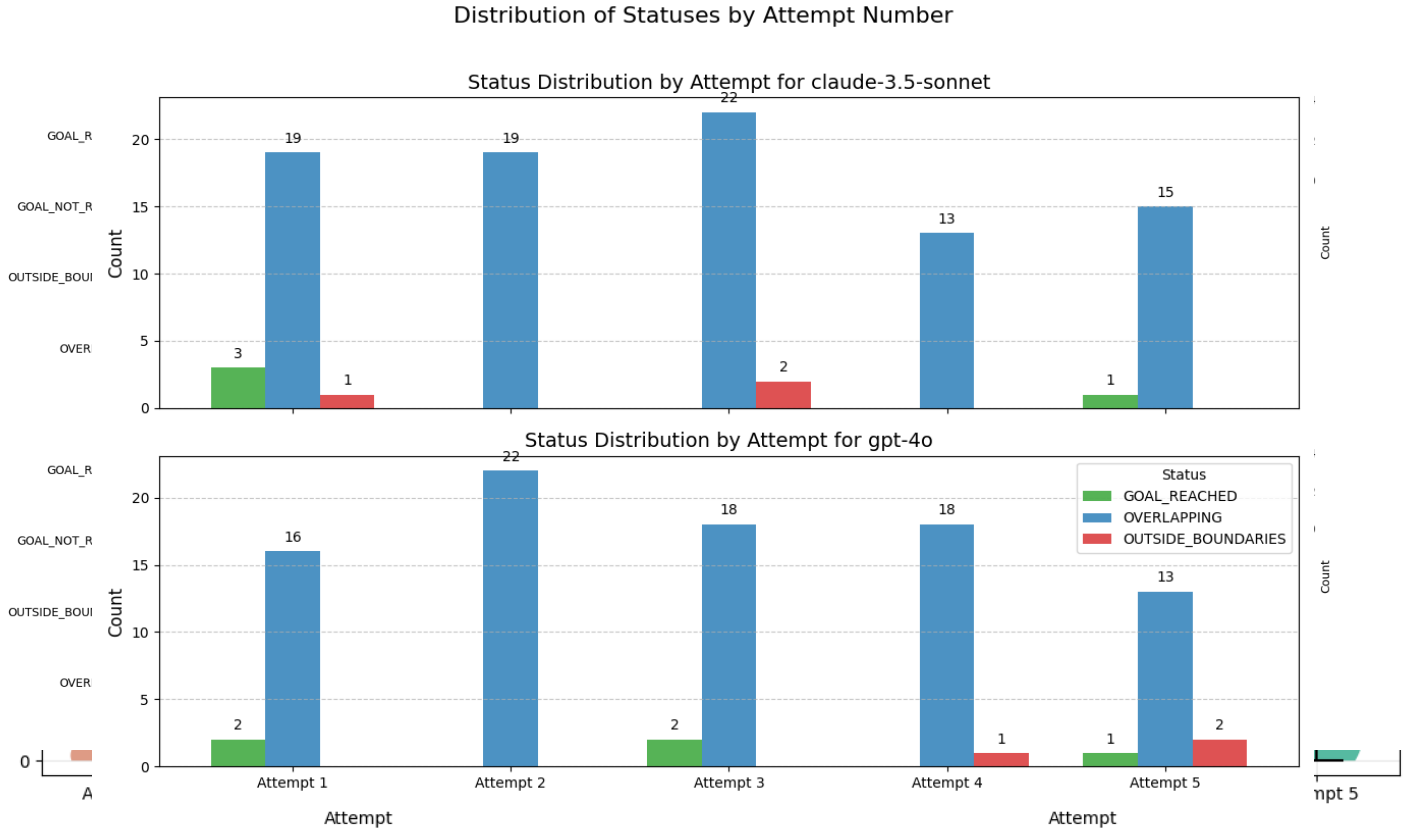
Interactive Insertion



Properties:

- # samples: 3 samples per template (150 total)
- No few-shot examples

Goal: determine weather the model can ascertain if the target objects are in contact



Conclusions & Future Work

Main results

- **Baseline Physical Perception**
  - The more advanced models demonstrated basic competence in identifying whether objects are in contact
- **Confidence & Calibration**
  - Confidence estimation weakened with increasing puzzle complexity
- **Binary Classification & Biases**
  - Strong biases and significant improvement for Claude and GPT-4o with few-shot examples
- **Ranking & Comparative Reasoning**
  - All models struggled, and adding examples didn't improve the performance
- **Interactive Evaluation**
  - Very low success rate, with frequent spatial errors (overlapping)

🔑 Takeaway: Current VLMs rely largely on learned correlations and struggle significantly with deep, iterative physical reasoning tasks

Future work

1. More Diverse Goals
2. Extend to 3D Environments
3. Real-Time Interaction & Extended Trials
4. Hybrid Approaches
5. Enhanced Prompting & Evaluation
6. Benchmark Expansion
7. Comparisons with Human Performance
8. ...

🎯 **Goal:** To create a robust, comprehensive benchmark that effectively measures true causal understanding and intuitive physics capabilities in AI systems.

Thank you

Made under the supervision of:

- Raquel Fernández (UvA)
- Alberto Testoni (UvA)
- David Shlangen (University of Postdam)
- Jacopo Staiano (UniTN)
- Roberto Dessì (FBK)