Séminaire IA - Université Gustave Eiffel

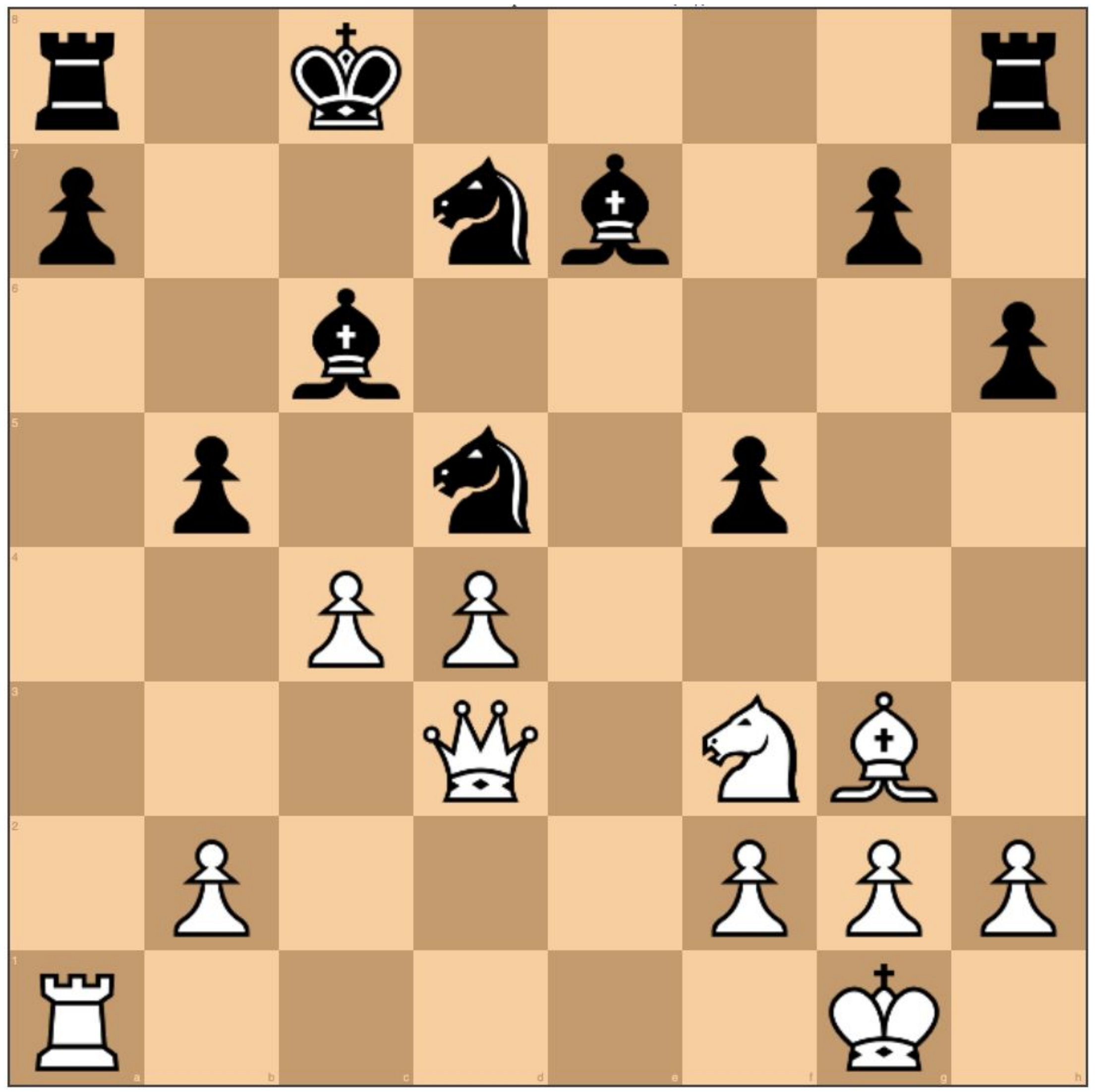
Learning to understand intuitive physics from natural videos

Quentin Garrido
FAIR at Meta, Université Gustave Eiffel - LIGM

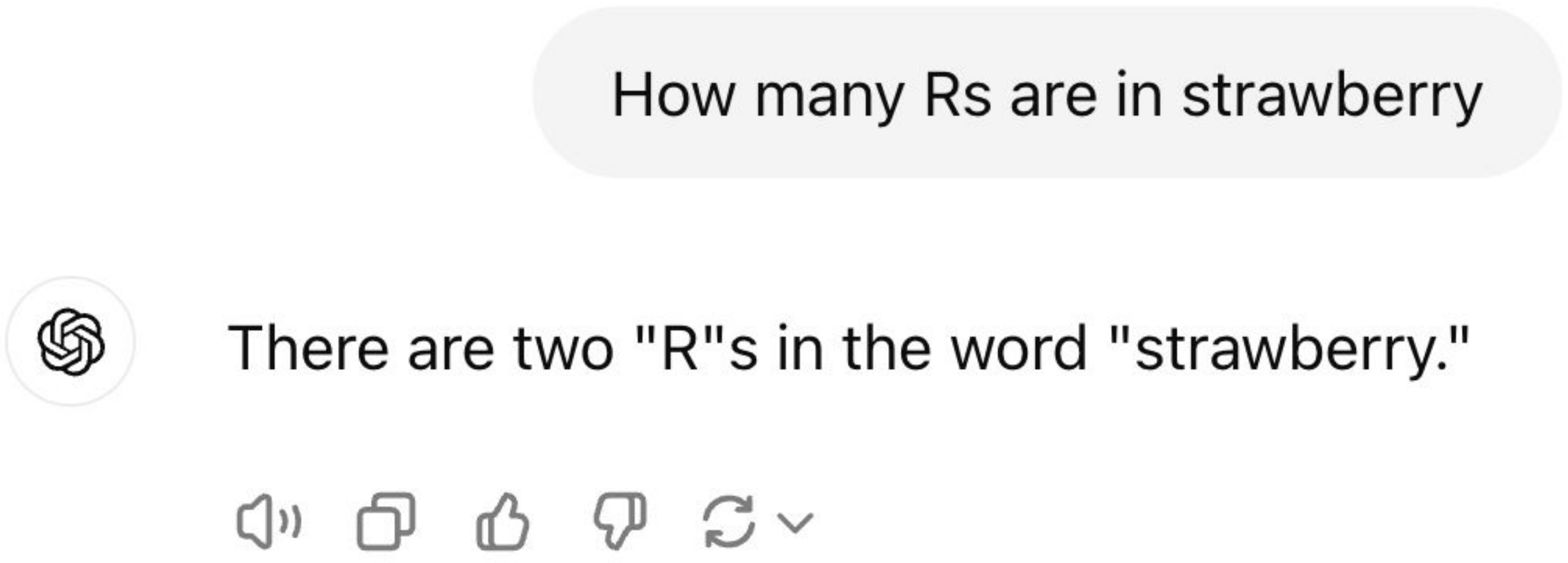
Joint work with: Mido Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, Emmanuel Dupoux, Mike Rabbat, Yann LeCun

We are building embodiments of Moravec’s paradox

Models are great at solving complex tasks…



… But they are terrible at tasks that are easy for us

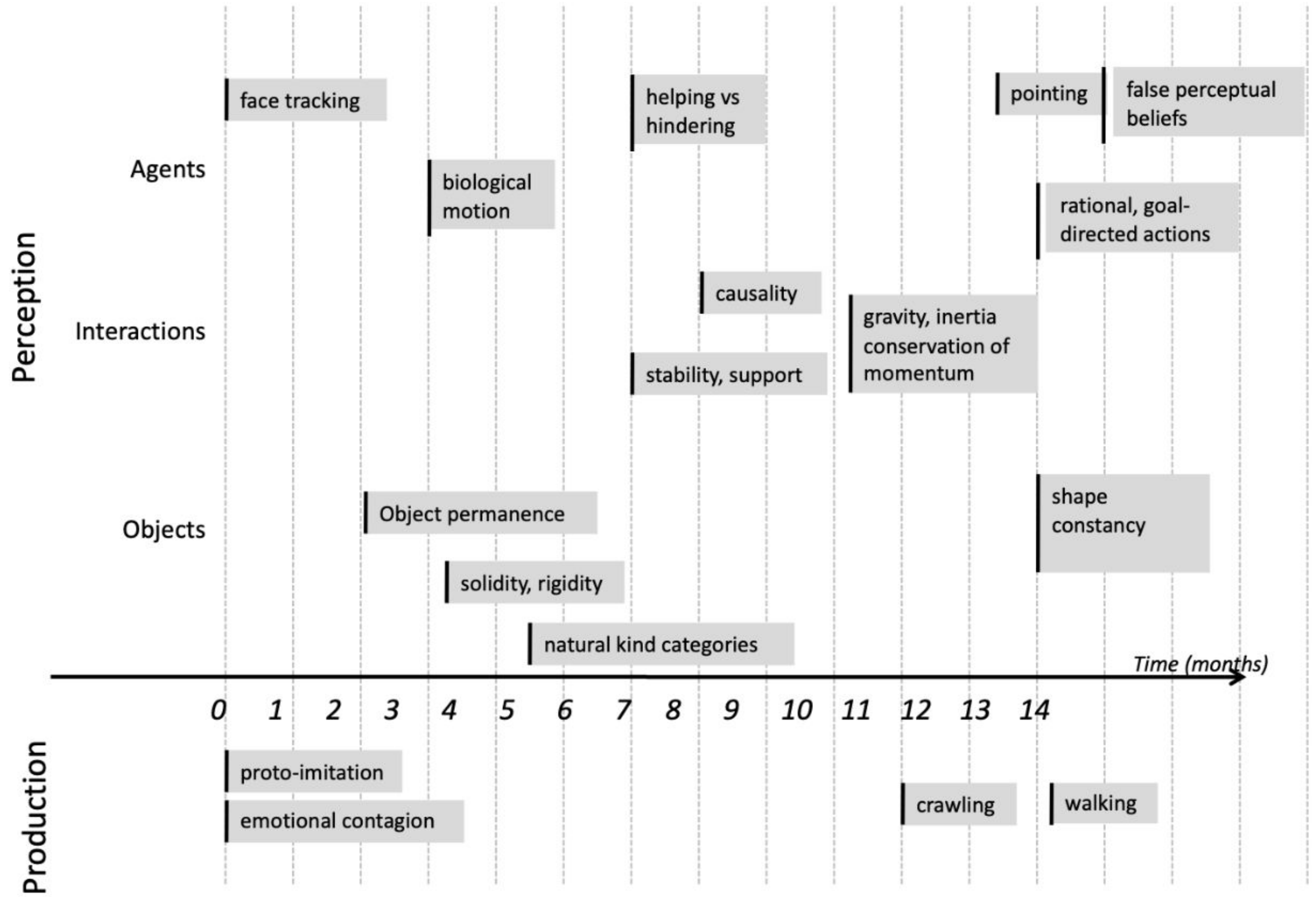


We should focus on learning like humans

Humans acquire an understanding of the physical world at a young age:

- Object permanence (objects don't appear out of nowhere) at 3 months
- Shape constancy (objects don't change shape suddenly) at 14 months
- Walking at 14 months
- Etc

Can we build machines that acquires these almost innate concepts, and break Moravec's paradox ?



Acquisition of intuitive physics properties in children.
From : https://arxiv.org/pdf/1803.07616

Intuitive physics benchmarks

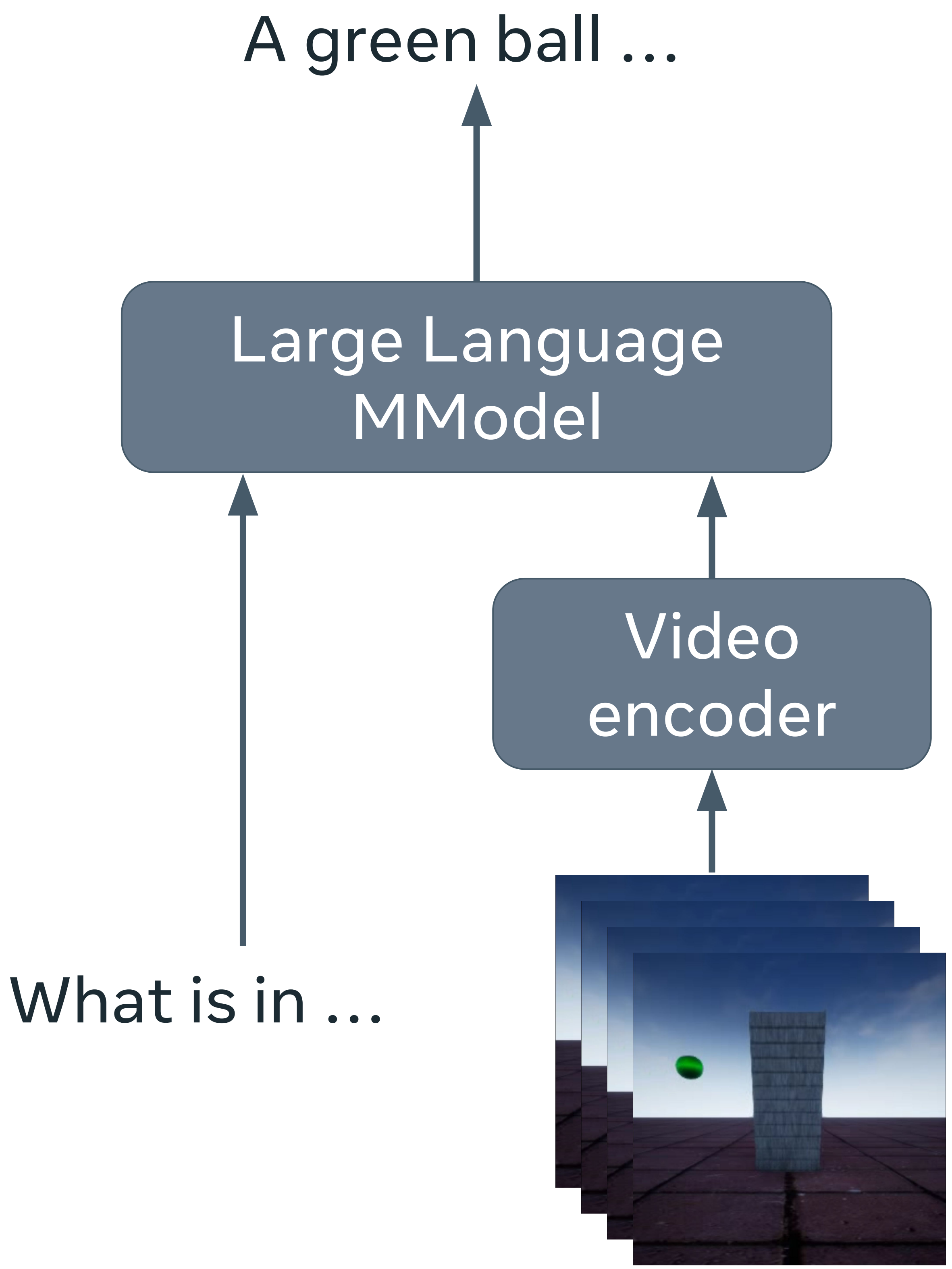
Can a model find which video is impossible ?
I.e. can a model detect violations of intuitive physical laws

Existing architecture: Multimodal LLMs

- Great at recognition tasks
- Give a textual interface

But …

- Need huge amount of compute/data
- Video support is often an afterthought

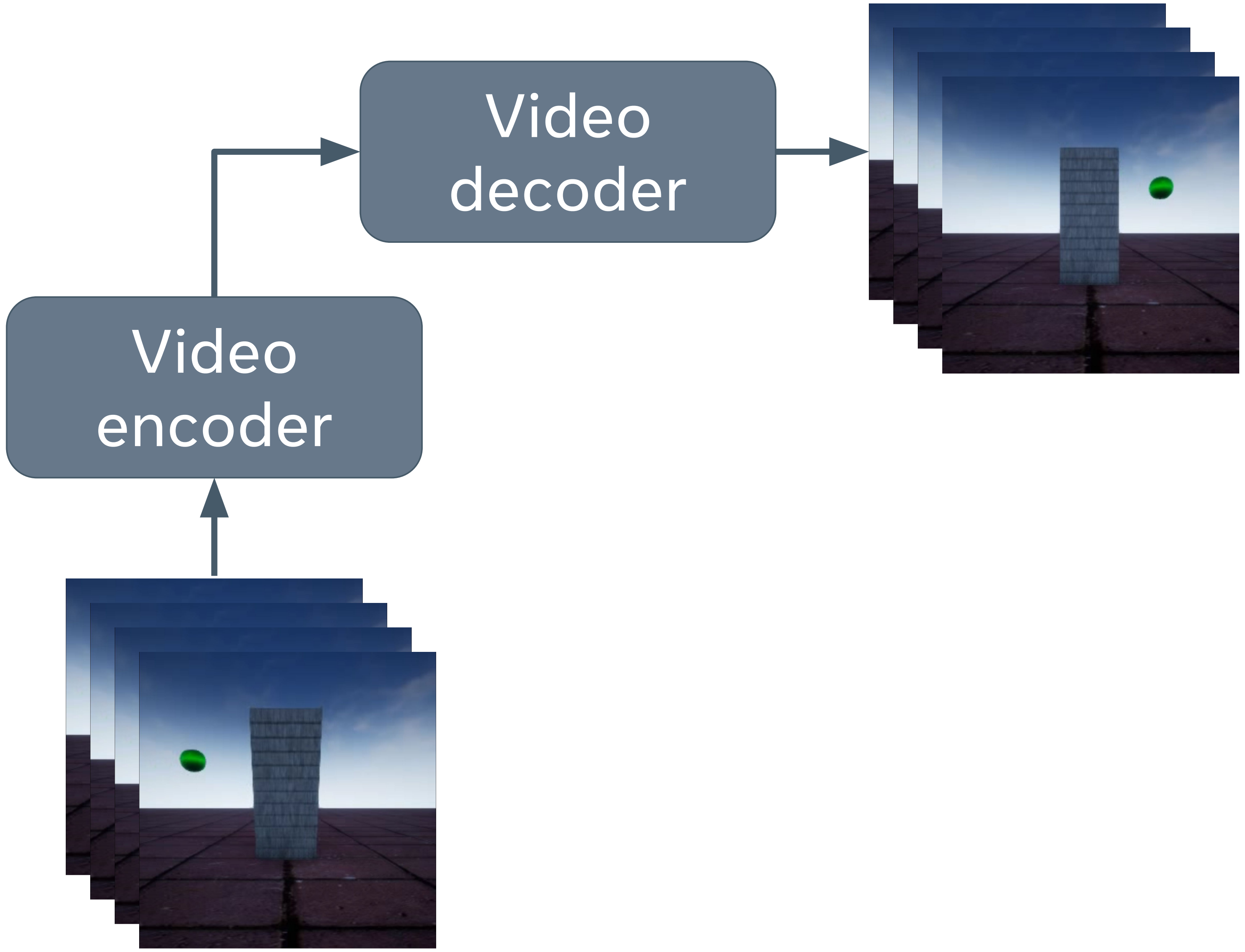


Existing architecture: Pixel Space Prediction

- Trained to predict the future
- We can look at predictions

But …

- Aren’t great at recognition tasks
- Focus too much on unnecessary details



These intuitive physics concepts are HARD

"The computational system generally performed poorly compared to humans"

IntPhys: Riochet et. al. (2019)

"Existing models struggle to identify violations of these principles despite their ability to accomplish other complex tasks"

InfLevel: Weihs et. al. (2022)

"the models generally exhibit performance equivalent to, or less than, chance across all tests"

GRASP: Jassim et. al. (2024)

Our goal:

Can we to build deep learning systems that learn like humans, exhibiting human-like understanding of intuitive physics ?

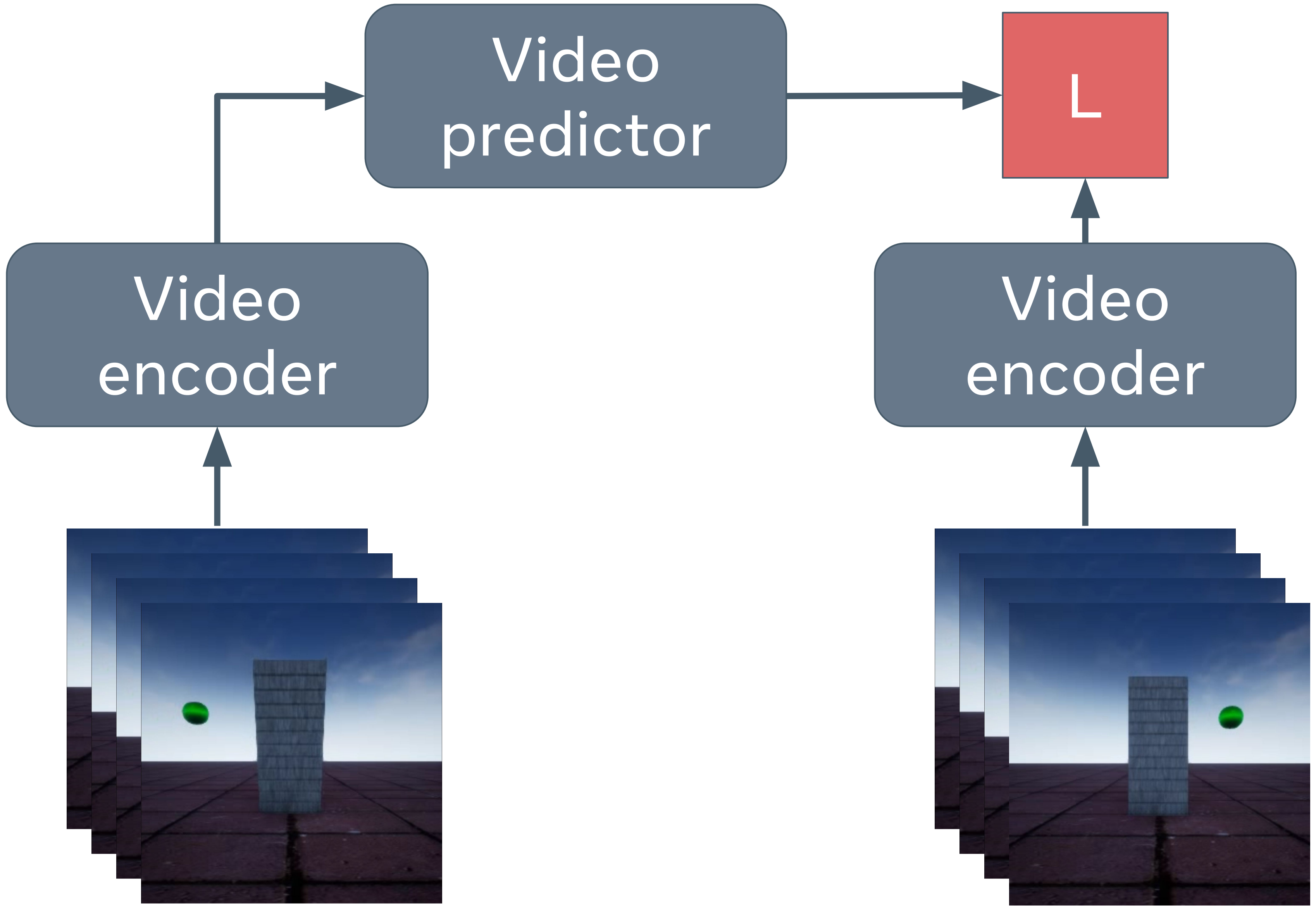
What we believe should be done: Latent Space Prediction

- Trained to predict in an abstract space
  - Predict the future
  - Inpainting
  - etc

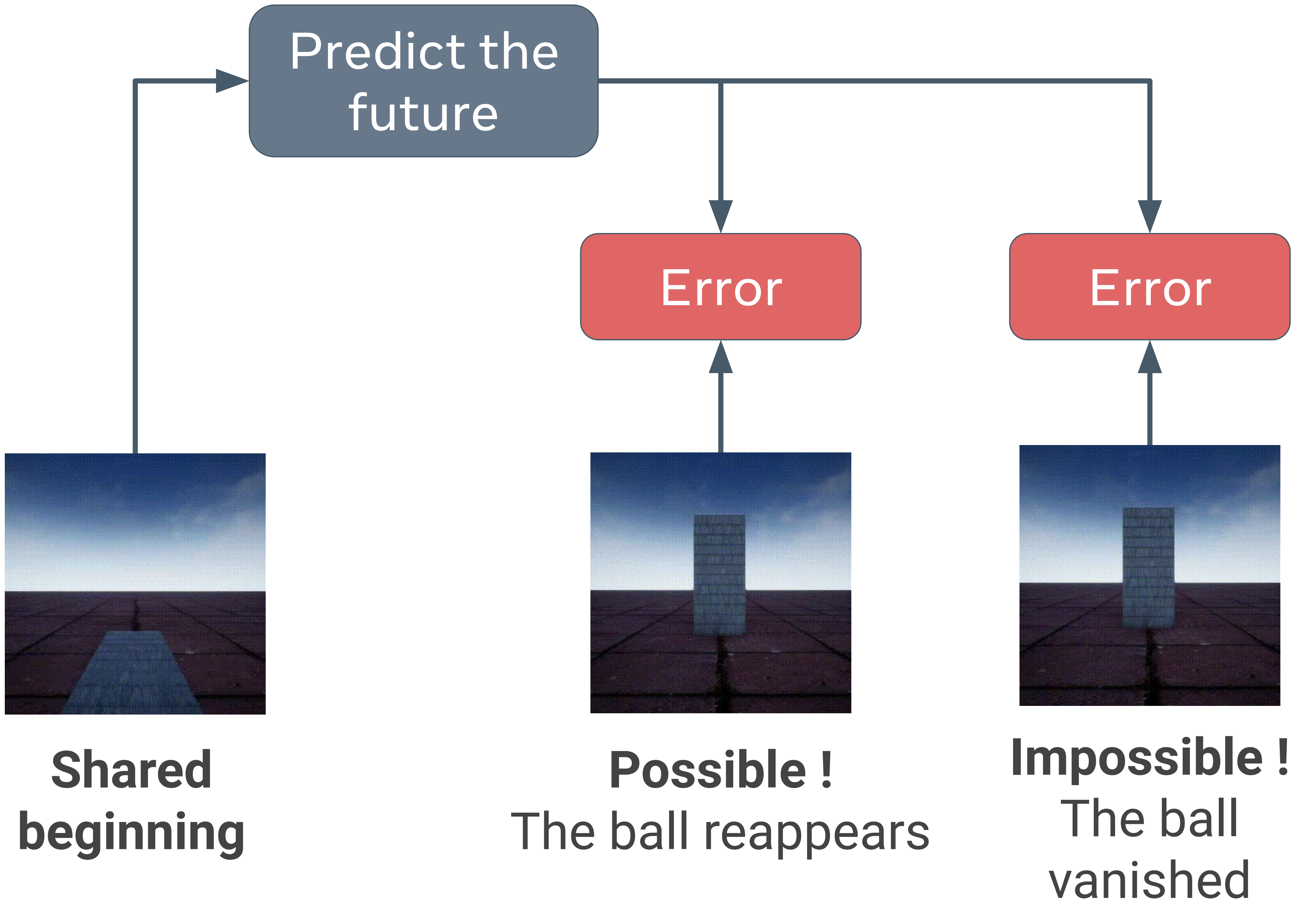
- Great at recognition tasks

But …

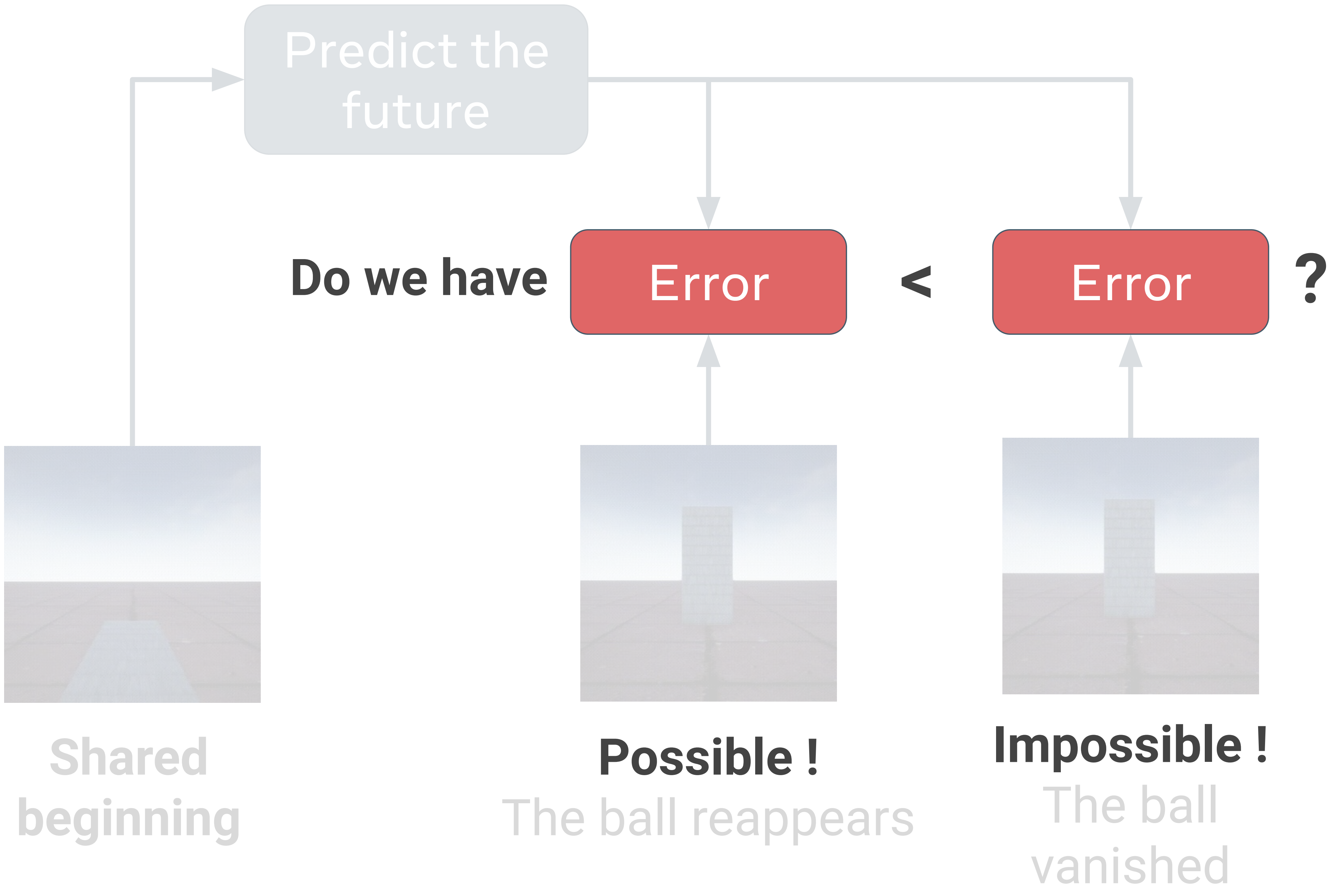
- No way to directly interpret predictions



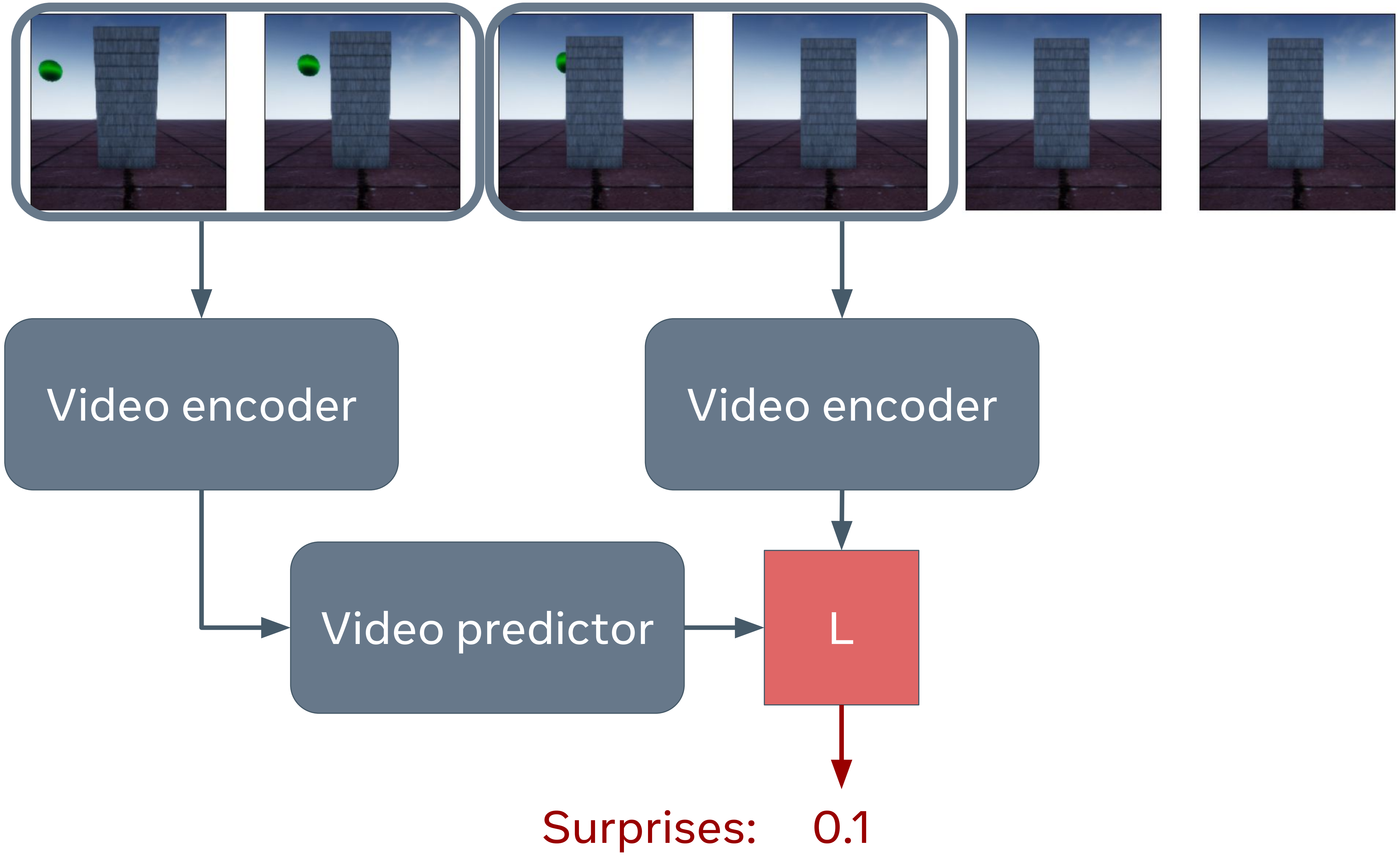
How to measure these properties: Violation of Expectation



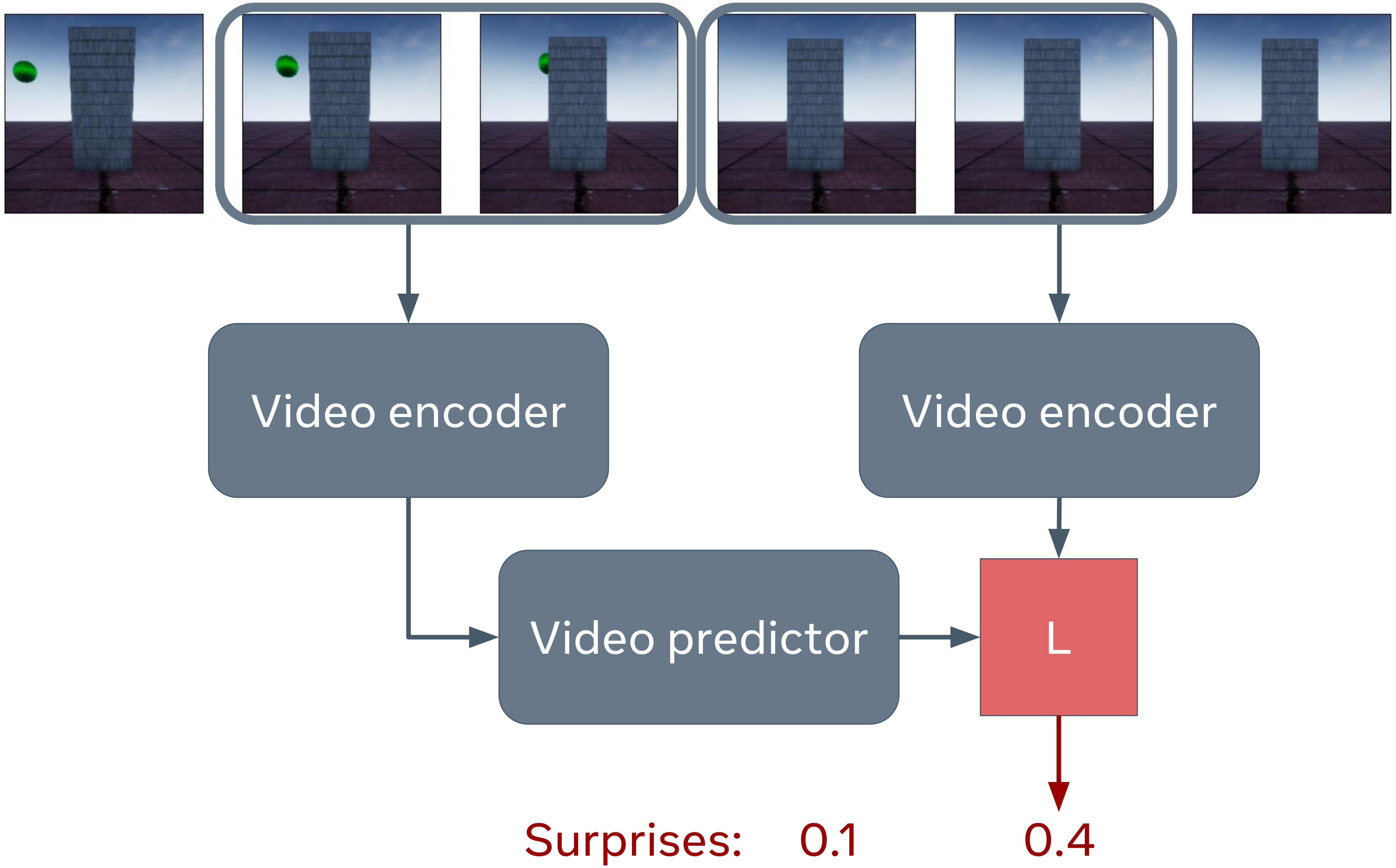
How to measure these properties: Violation of Expectation



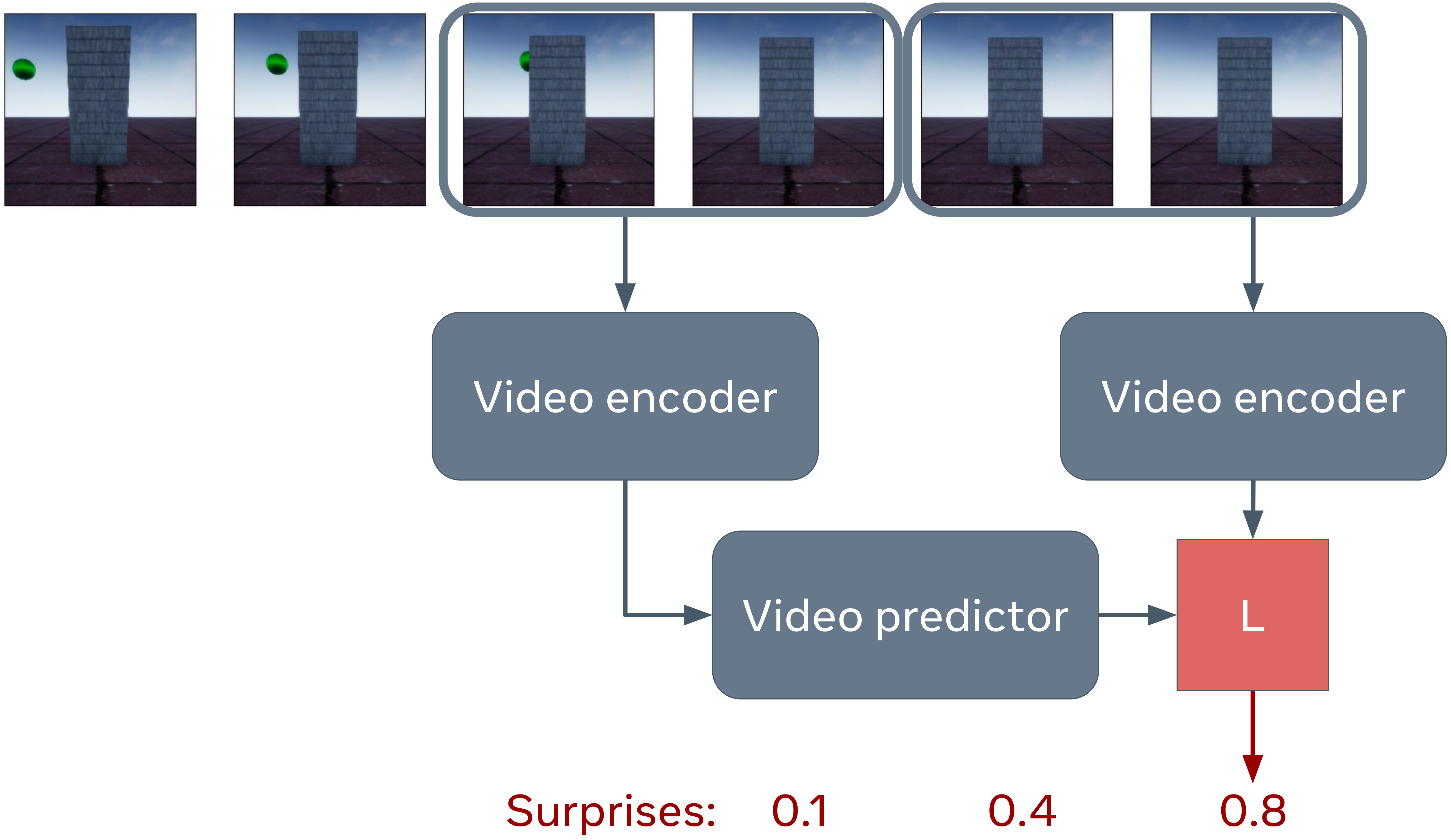
Violation of Expectation over time



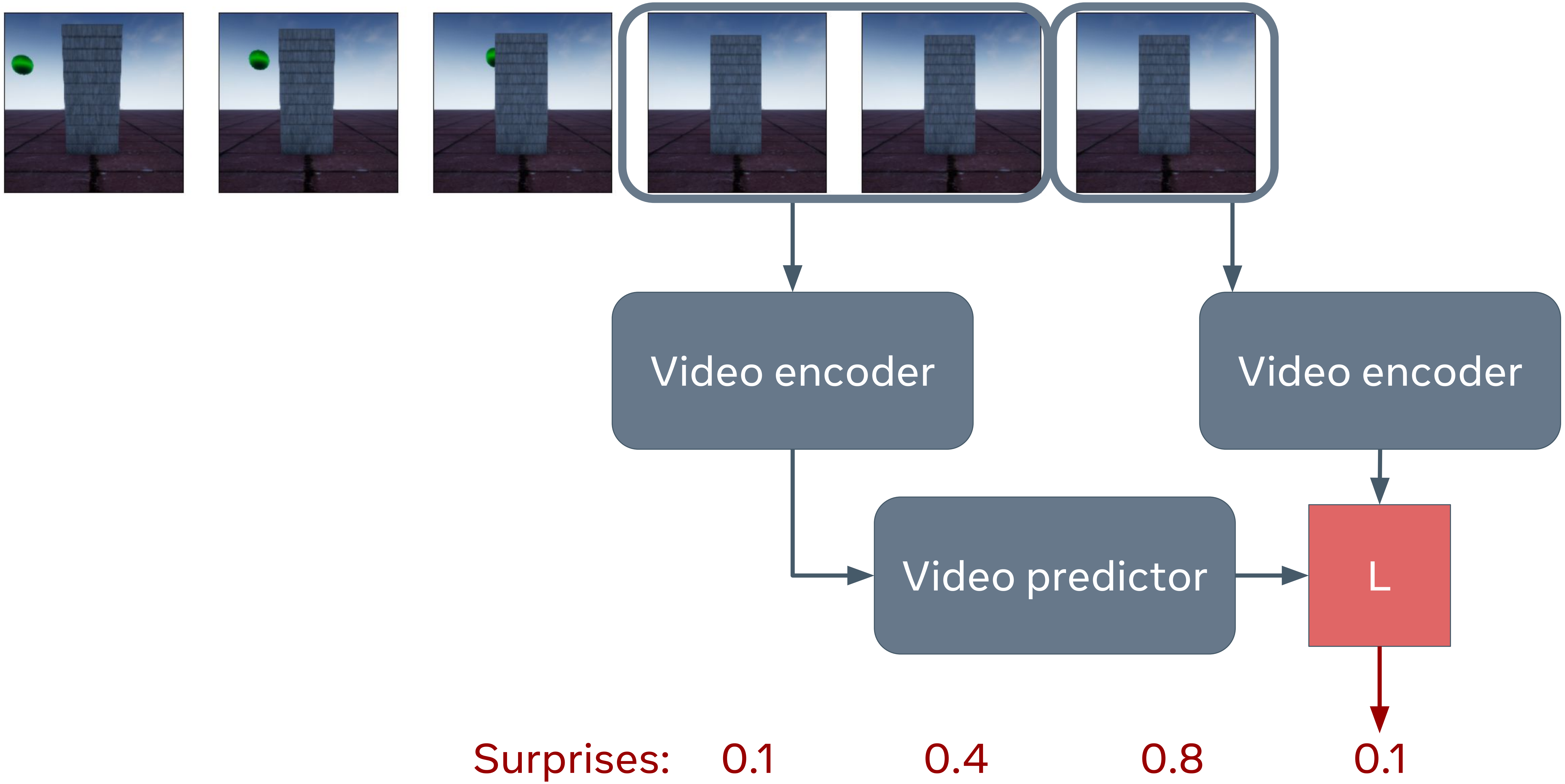
Violation of Expectation over time



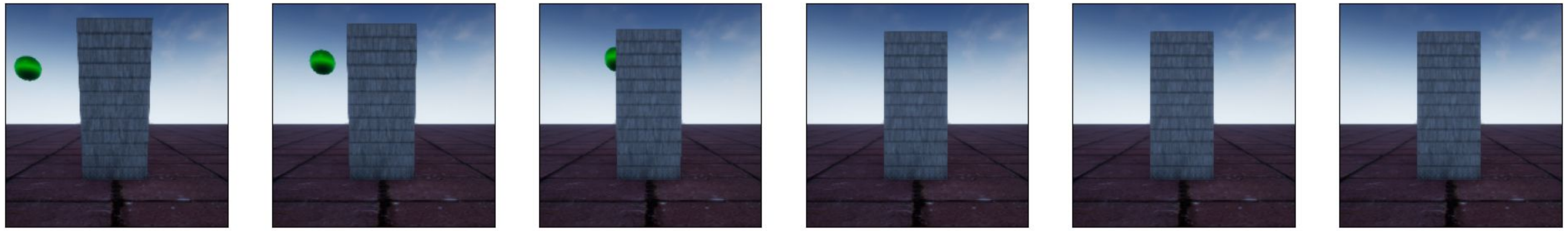
Violation of Expectation over time

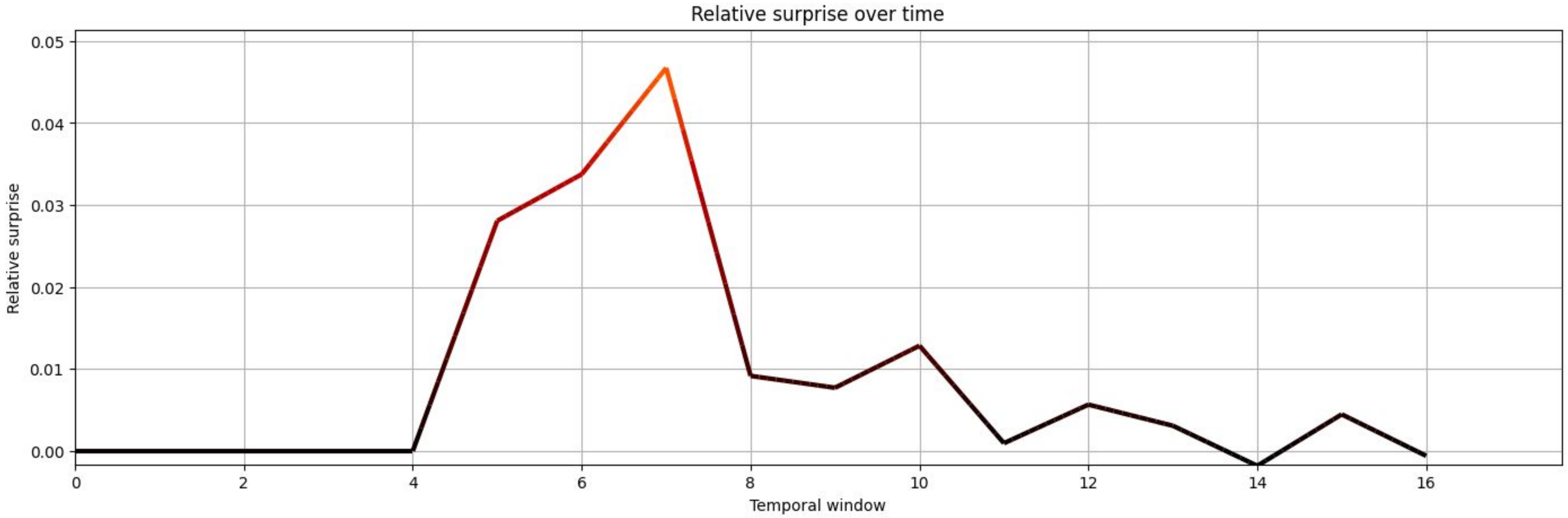


Violation of Expectation over time



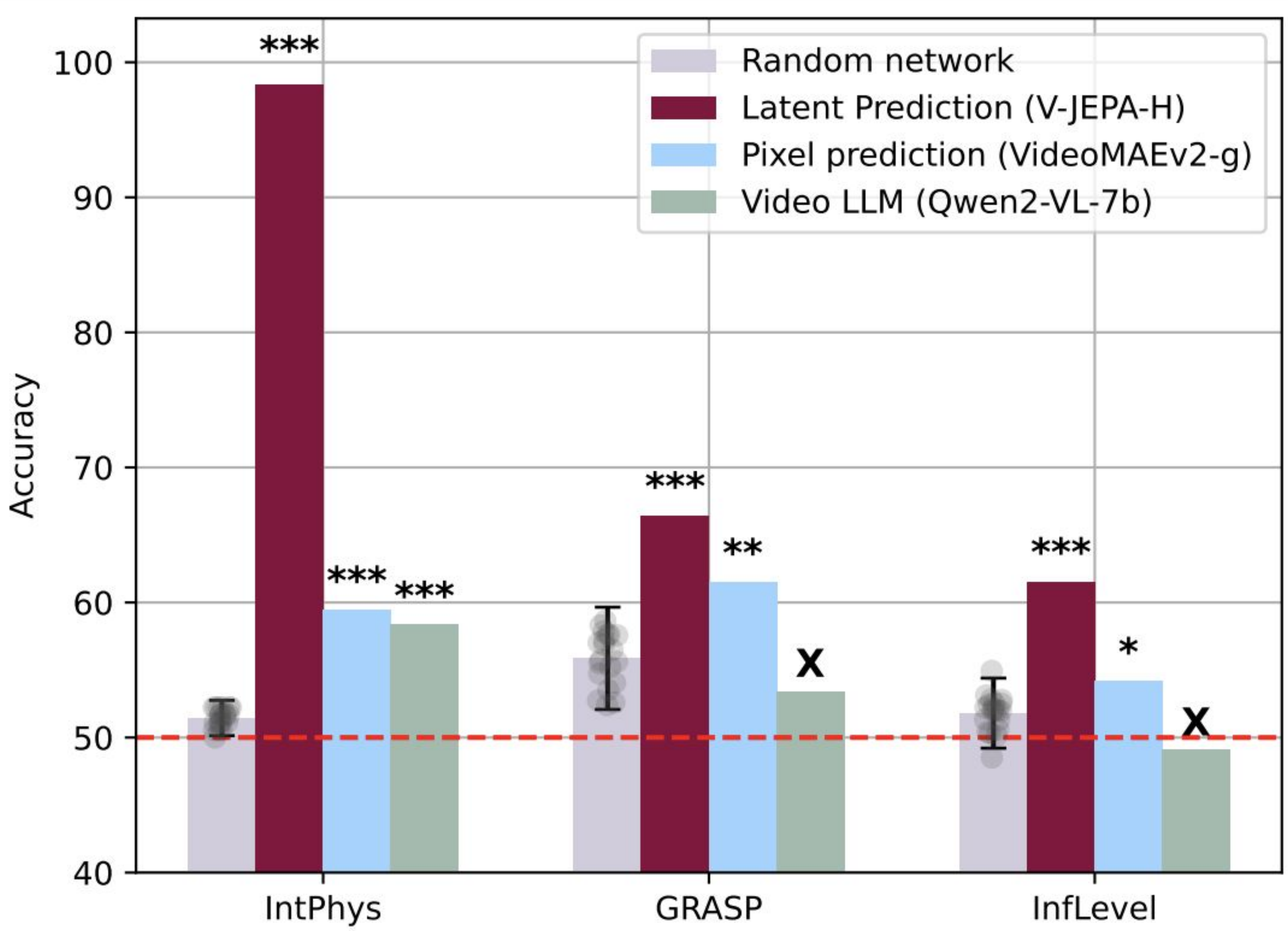
Violation of Expectation over time



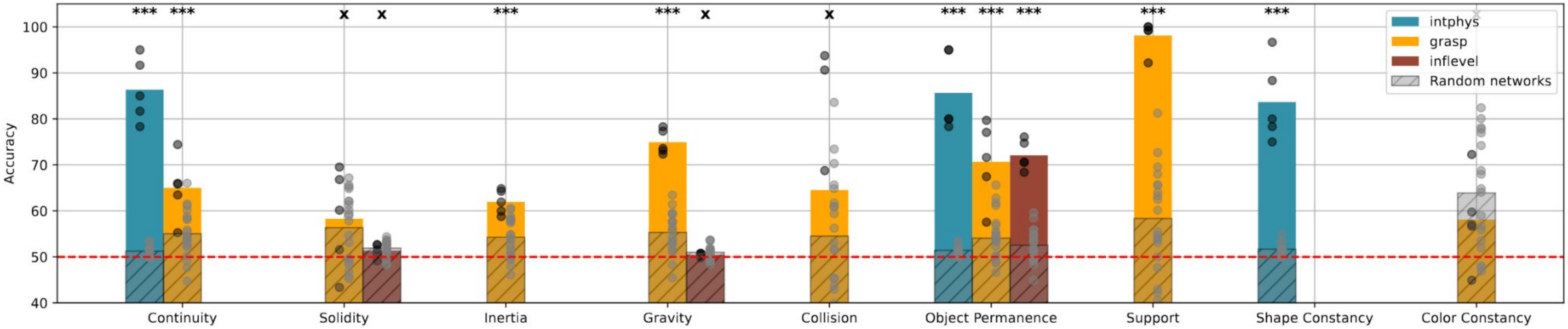


The model should show a higher surprise when something physics breaking happens

Only Latent Prediction can reliably find the impossible video

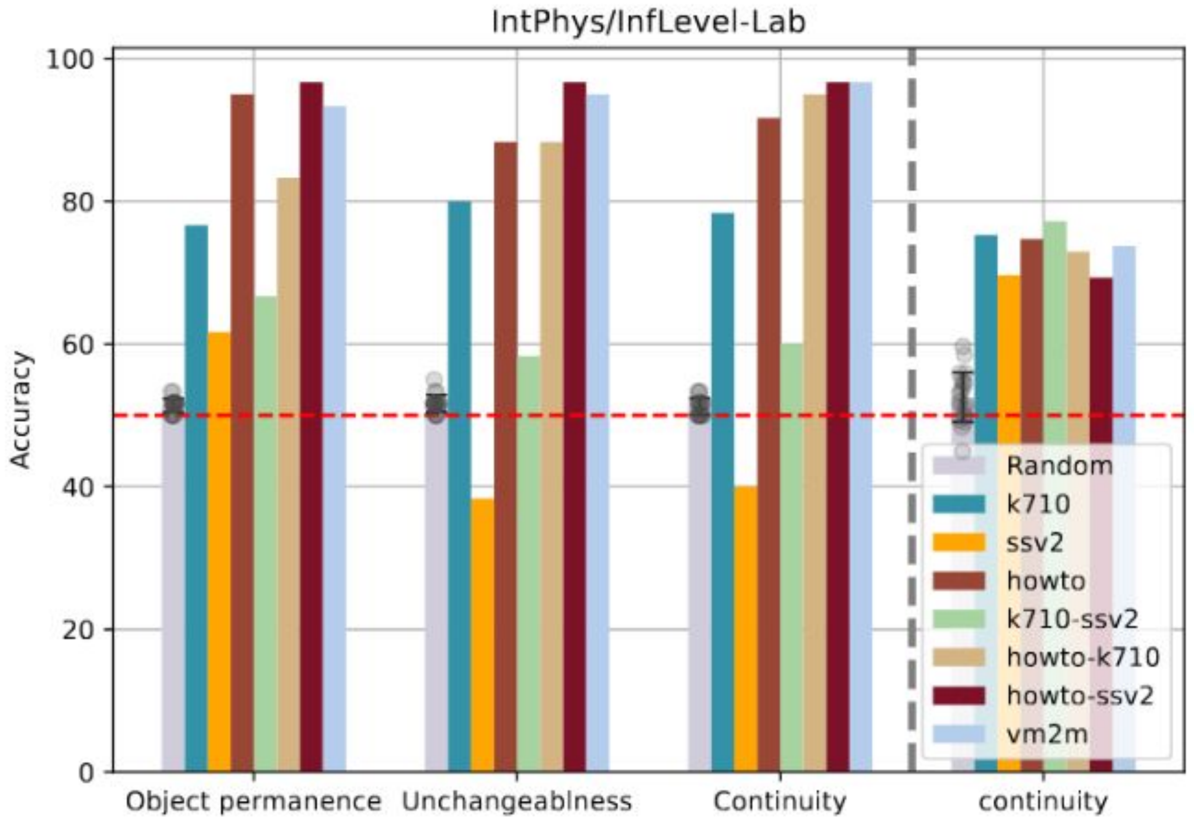


We find a non trivial understanding across properties



We find non significant performance when too much memory is requires, or when the task is flawed (a random network can sometimes perform well)

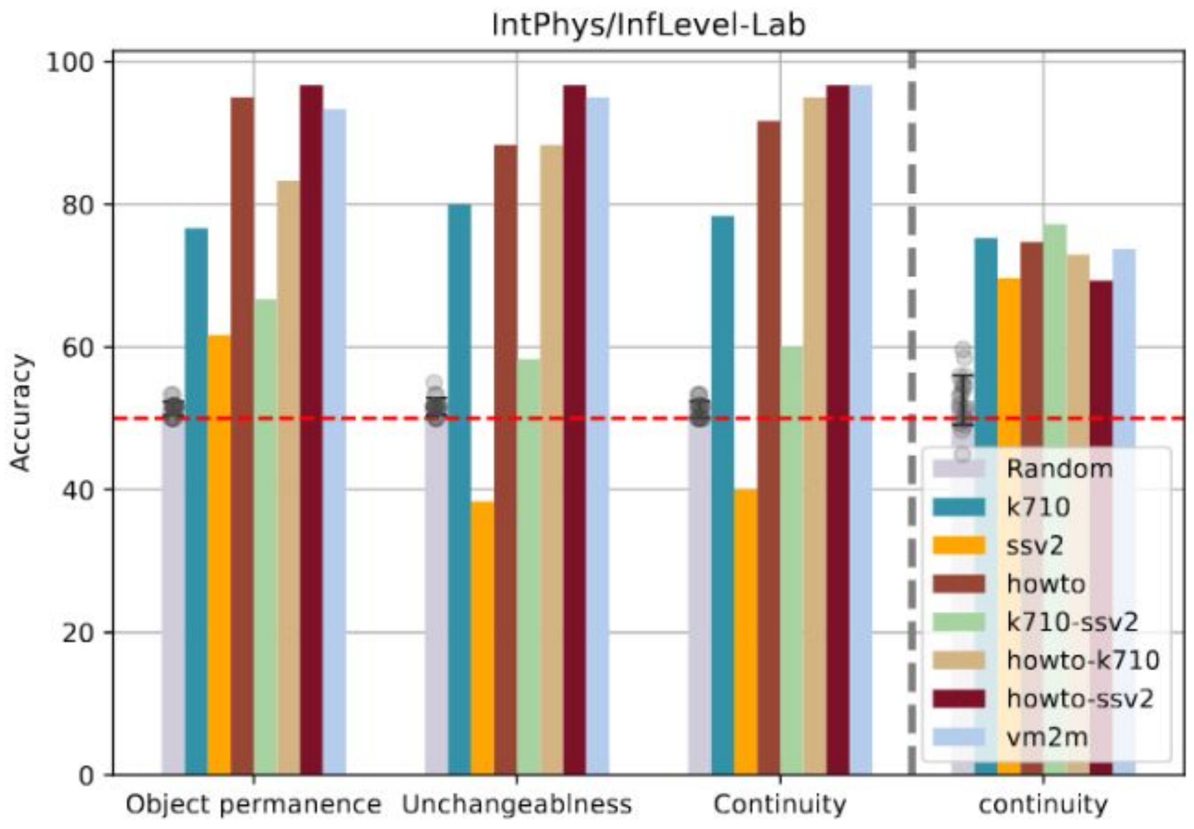
Why does this understanding emerge ? The right data



Some data sources are better than others.

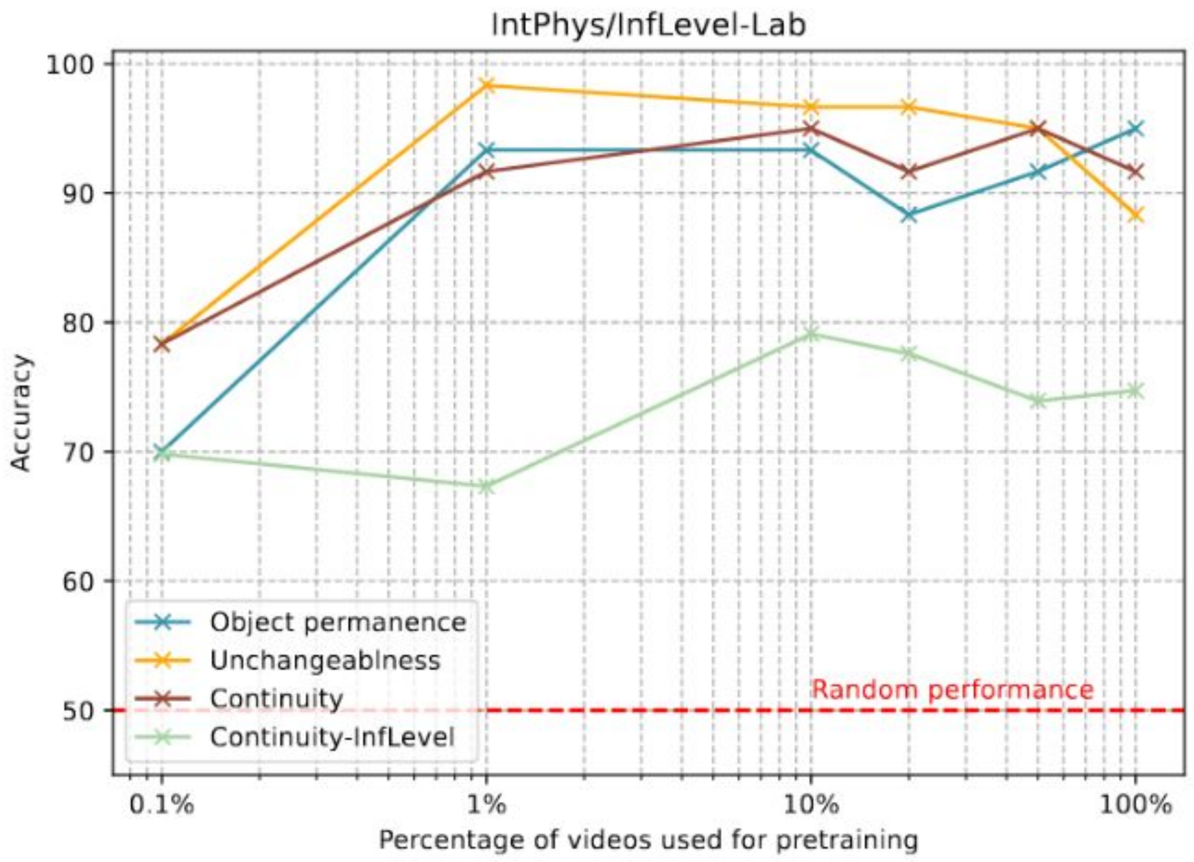
Youtube tutorials are ideal.

Why does this understanding emerge ? The right data



Some data sources are better than others.

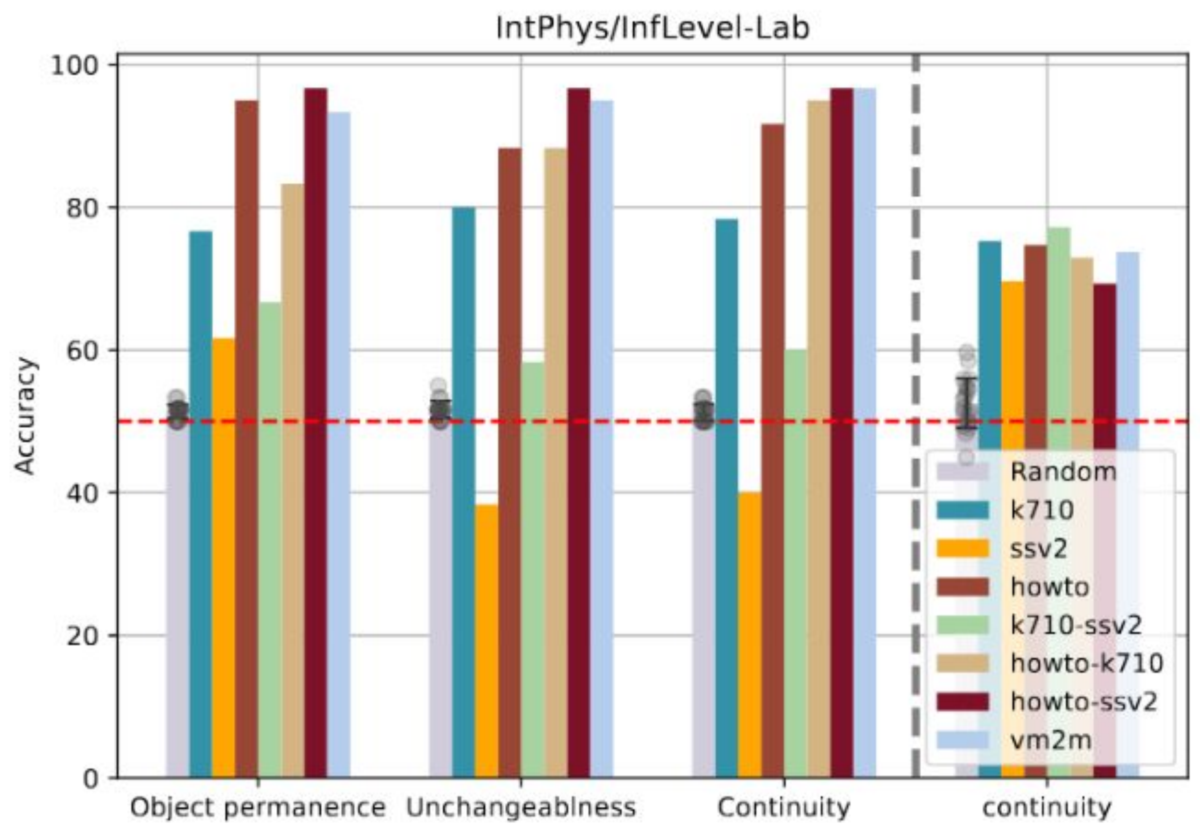
Youtube tutorials are ideal.

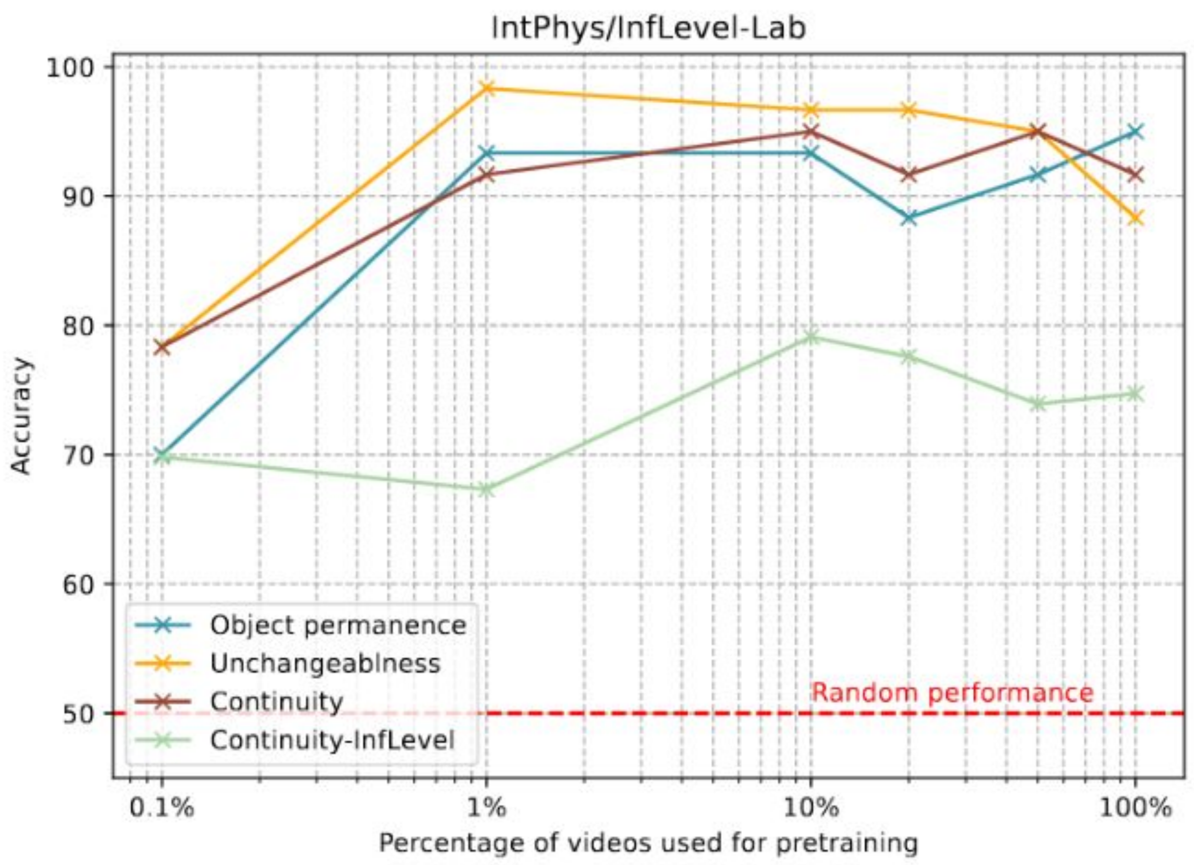


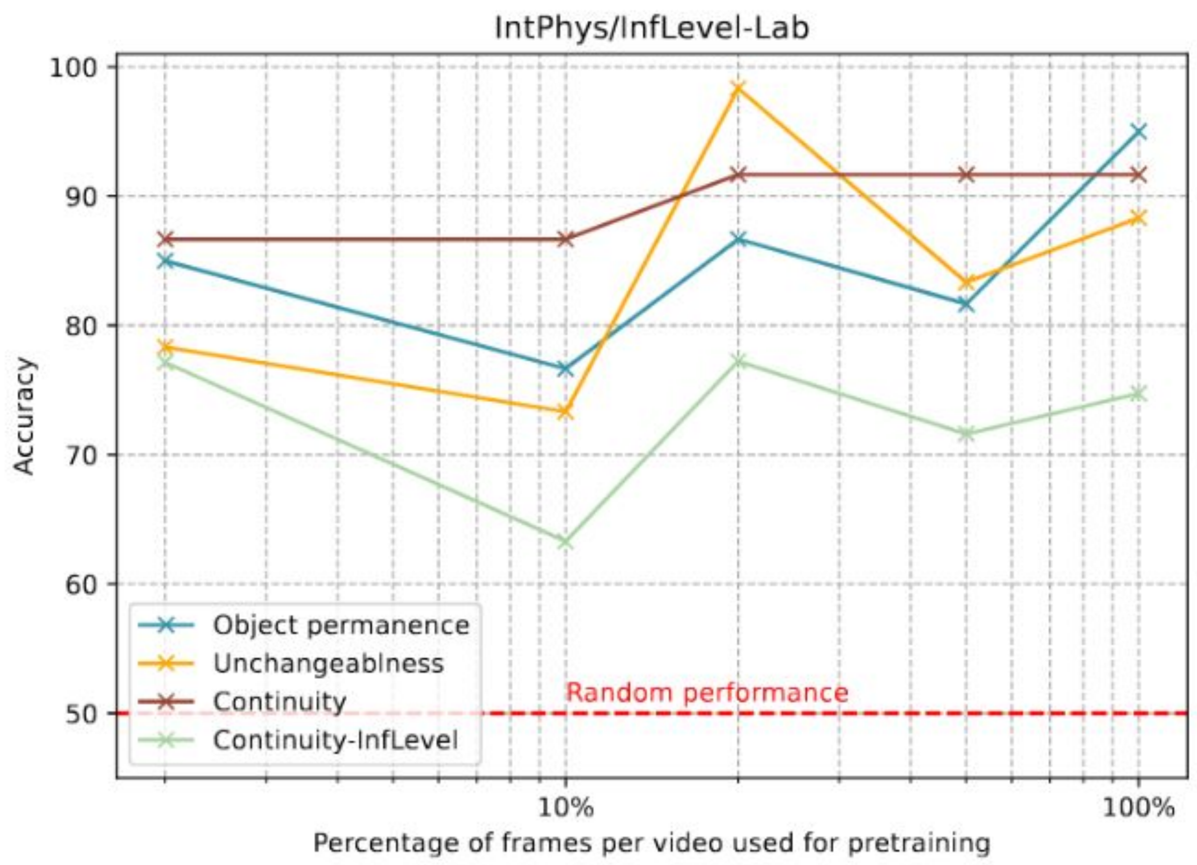
We only need a little bit of unique videos.

As little as 150h is enough.

Why does this understanding emerge ? The right data







Some data sources are better than others.

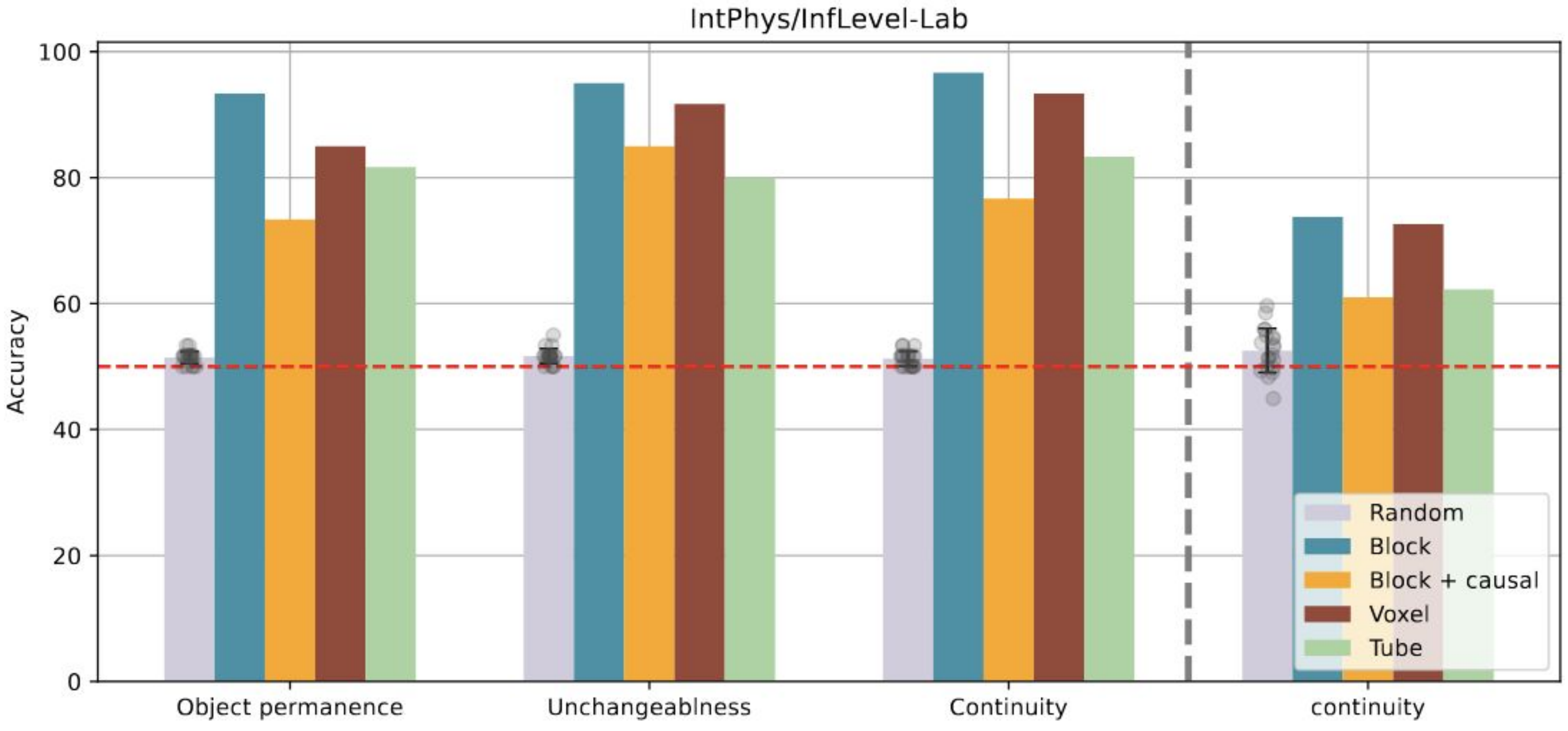
Youtube tutorials are ideal.

We only need a little bit of unique videos.

As little as 150h is enough.

It's better to see fewer videos than shorter ones.

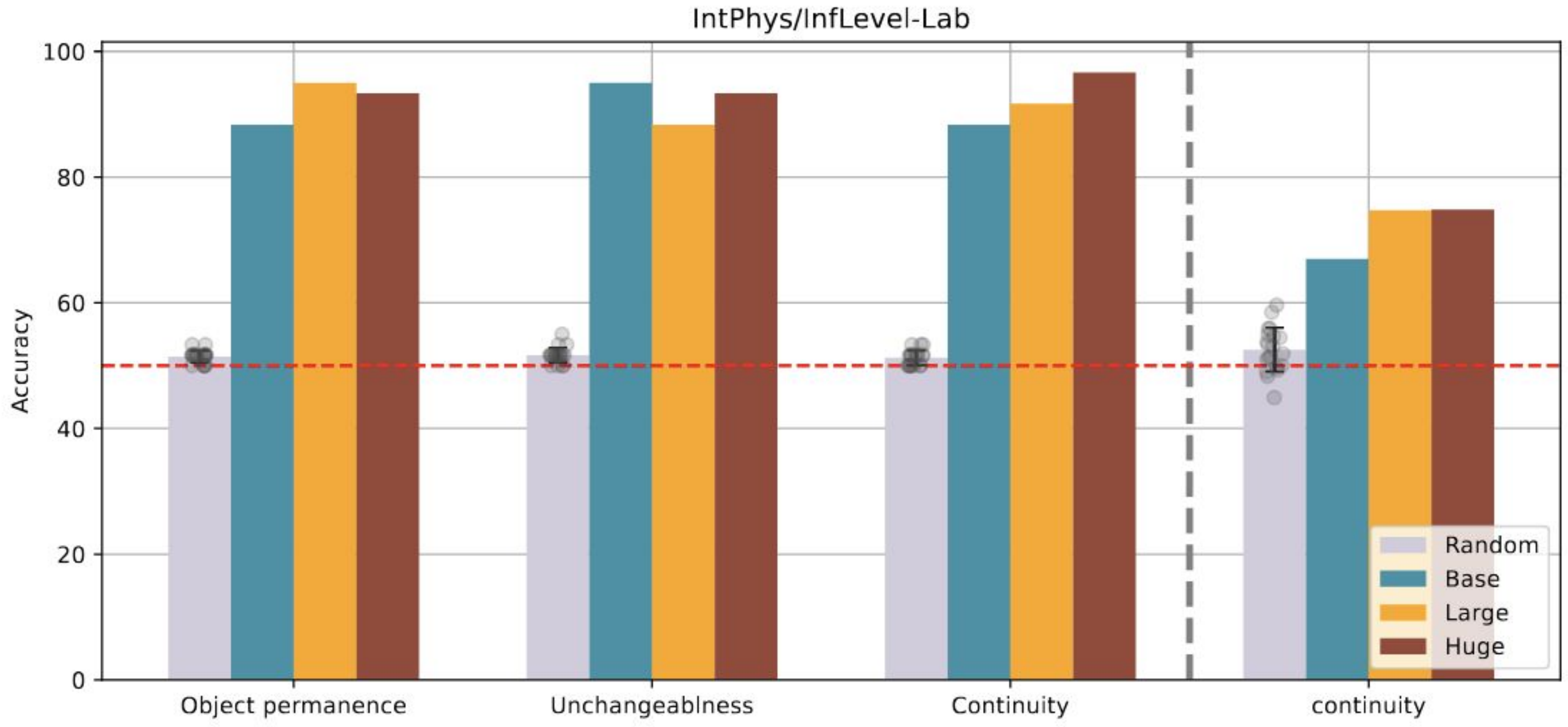
Why does this understanding emerge ? The right pretraining



Different pretraining prediction tasks change performance.

The key is to predict in latent space.

Why does this understanding emerge ? The right size



Bigger models are better, but 80M parameters is enough.

Compare that to 10s of billions for LLMs...

How close is our paradigm to babies

How do babies learn ?

They observe the world

They interact with the world

They don't speak/read

They are young

How to translate to neural networks

How close is our paradigm to babies

How do babies learn ?

They observe the world → Use natural videos from the wild

They interact with the world

They don't speak/read

They are young

How to translate to neural networks

How close is our paradigm to babies

How do babies learn ?		**How to translate to neural networks**
They observe the world	→	Use natural videos from the wild
They interact with the world	→	Can predict the consequences of actions
They don't speak/read		
They are young		

How close is our paradigm to babies

How do babies learn ?		How to translate to neural networks
They observe the world	→	Use natural videos from the wild
They interact with the world	→	Can predict the consequences of actions
They don’t speak/read	→	No need for text data
They are young		

How close is our paradigm to babies

How do babies learn ?		How to translate to neural networks
They observe the world	→	**Use natural videos from the wild**
They interact with the world	→	**Can predict the consequences of actions**
They don’t speak/read	→	**No need for text data**
They are young	→	**Should not need gargantuan amounts of data**

How close is our paradigm to babies

How do babies learn ?		How to translate to neural networks
They observe the world	→	Use natural videos from the wild
They interact with the world	→	Can predict the consequences of actions
They don't speak/read	→	No need for text data
They are young	→	Should not need gargantuan amounts of data

Most importantly:

Latent Prediction is the only framework that understands intuitive physical concepts non trivially.

Take home messages

- Current deep learning approaches are very non-human in their learning, failing at simple intuitive physics tasks
- We want methods that exhibit more similarities to humans
- Latent Video Prediction offers the first non trivial performance at intuitive physics benchmarks

We should scale frameworks that learn like humans rather than scaling approaches that don't and hoping that they will exhibit human behaviour at some point.