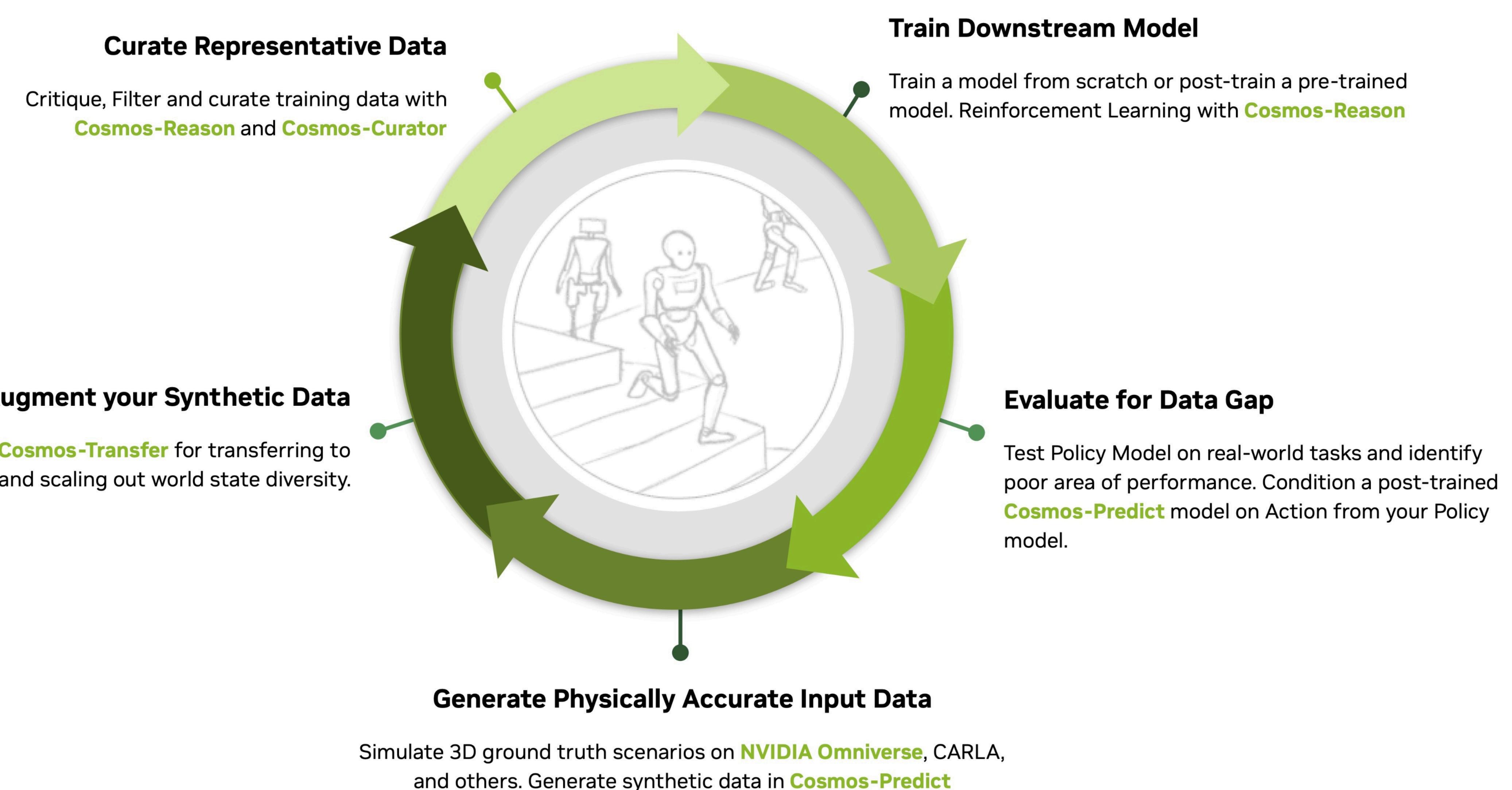


Cosmos-Predict

Cosmos-Transfer

World Simulation with Video Foundation Models for Physical AI

Physical AI Data Flywheel



Strength

Training data domain

- deliberately excludes cartoons, games, animations
- filtered to physical-world video only
- Includes specialized robotics datasets (DROID, GR0OT, etc.)

Domain-specific SFT

- separate fine-tuned models for object permanence, high motion, complex scenes, driving, robotic manipulation then merged

Efficiency results

- 2B model was matching Wan2.2-27B on PAI-Bench

Relevant capabilities

- Sim2Real
- Multi-view generation: synthesize 3 synchronized camera views from a single input view
- Text-controlled scene variation (objects, lighting, backgrounds)
- Action-conditioned prediction: generate video given current image and 7-DoF actions
- Long-video generation: autoregressive chunking with reduced error accumulation vs. predecessor

Benchmarking

- Benchmarked primarily on PAI-Bench, conflict of interests?
- Not on Text2Video arenas due to specificity

Ablations and Experiments

- No ablations isolating which ingredient matter most (flow-matching, RL, etc.)
- Robot experiment is limited to 1 task, 100 demos, 3 trials per condition

Weaknesses

Other relevant to challenge

- No/limited deforming objects
- Physical plausibility dataset curated but never evaluated
 - VideoPhy 2
 - PhyGenBench
- action-condition model and multiview model are separate, were not combined into one
 - Intrinsic has 3 wrist cameras, a policy evaluator would need: planned actions and what 3 cameras would see -
 - two outputs most won't be consistent with each other because they weren't generated jointly

- Take robot's practice runs in the simulator and make them look real before using them to teach the robot.

Strongest Use Case

- Tested on a robot: trained with this approach, it succeeded 24 out of 30 times across 10 different unexpected scenarios. Without it: 1 out of 30.