Critical Review

Kris Carlson prompting ChatGPT DeepResearch

Summarize this paper noting especially novel views of common sense, its definition and theory of common sense, and how a connection to the physical world informs its theory, its common sense benchmarks, and then critique its strengths and weaknesses. https://arxiv.org/abs/2503.15558

NVIDIA: Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning

From Physical Common Sense To Embodied Reasoning is an arXiv paper (36 pp., by a team at NVIDIA) that proposes a framework for endowing AI systems with common sense grounded in the physical world. The authors introduce a theoretical **definition of common sense** in terms of key capabilities, connect this theory to physical embodiment (perception and action), and implement it in **Cosmos-Reason1**, a family of multimodal large-language models. They also design **benchmarks** to evaluate physical common sense and embodied reasoning. Below, we summarize the paper's main contributions and analyze its strengths and weaknesses.

Defining Common Sense: Ontologies and Novel Perspectives

The authors define **physical common sense** as a general, embodiment-agnostic understanding of how the physical world works ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). In other words, an AI with common sense should know what is plausible or implausible in reality, independent of any specific robot or human body. Humans acquire such knowledge early (e.g. infants learn object permanence and basic gravity within months ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning)), and it allows us to predict outcomes and avoid dangerous mistakes. The paper argues that equipping AI with this kind of knowledge will let it learn new skills with minimal trial-and-error and operate more safely in novel situations ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning).

To formalize what "common sense" entails, the authors introduce a **hierarchical ontology** of physical knowledge. It has three top-level categories – **Space**, **Time**, and **Fundamental Physics** – further broken down into 16 specific sub-categories ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). This ontology focuses on capabilities rather than any particular internal representations or algorithms, which is a novel perspective (inspired by Morris et al. 2024) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). In other words, it lists what an AI should be able to understand (e.g. "object persistence" or "temporal order of events"), without prescribing how the AI must achieve it or what form the AI's body takes ([2503.15558] Cosmos-Reason1: From Physical Common Sense To

Embodied Reasoning). The authors emphasize that an AI can have common sense without necessarily behaving exactly like a human; for example, it should grasp spatial relations and physics even if it doesn't have a humanoid body ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). This broad view of common sense knowledge – decoupled from specific embodiment – is a notable aspect of their theory.

([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) Illustration of the paper's physical common sense ontology (Figure 2), dividing common sense knowledge into three main categories – **Space** (red), **Time** (green), and **Fundamental Physics** (blue) – with 16 fine-grained subcategories ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). For example, Space includes understanding object relationships, spatial plausibility and affordances; Time includes actions, event ordering, and causality; Fundamental Physics covers object attributes and states, mechanics, thermodynamics, and even detecting physically impossible situations ("Anti-Physics").

In addition to common sense reasoning about how the world works, the paper asserts that true common sense in AI requires **embodied reasoning** – the ability to use that knowledge to **perceive, plan, and act** in a physical environment ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). They distinguish this from abstract problem-solving: unlike solving a math equation, an embodied agent must deal with continuous sensory inputs, unforeseen changes, and the constraints of physics in real time ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). The authors draw on cognitive psychology, aiming to incorporate both **"System 1"** (fast, intuitive responses) and **"System 2"** (slow, deliberative reasoning) in their model of common sense ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). In practice, this means an AI should have immediate intuitive judgments about physical situations and the ability to reason through complex scenarios step by step. Emphasizing both systems is a novel perspective, aligning AI reasoning with human-like common sense cognition.

Embodiment, Perception, and Physical Interaction

A core claim of the paper is that common sense must be connected to the **physical world through embodiment**. The authors describe embodied reasoning as reasoning that is grounded in action: an AI agent not only interprets what it perceives, but also imagines or plans the **next actions** to take in a dynamic environment ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). They outline four key capabilities an embodied AI needs for this:

- **Processing complex sensory inputs:** The AI must handle raw, potentially noisy perceptual data (e.g. video frames) and extract meaningful patterns ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning), rather than relying on neat, symbolic inputs. This is crucial because real-world data is messy and incomplete.
- **Predicting action effects:** The AI should have an intuitive sense of cause and effect – if it or another agent performs an action, what will happen next? For instance, it should predict

how pushing an object might make it fall, or how a robot's movement will be affected by the terrain ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning).

- **Respecting physical constraints:** Any plan must obey physical laws like gravity, friction, object solidity, etc. The AI's reasoning should filter out actions that are physically impossible or dangerous (you can't lift a car with one hand; a wheeled robot can't instantly stop on ice) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). Long-horizon action sequences need to remain feasible at each step.
- **Learning from interaction:** As the agent acts and the world responds, the AI should update its understanding. This continuous learning loop allows adapting to new circumstances ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). (Notably, the authors leave this fourth aspect as future work, focusing their current study on the first three capabilities ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning).)

To systematically handle embodiment, the paper introduces a **two-dimensional ontology** for embodied reasoning ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). One dimension enumerates the reasoning capabilities above; the other dimension spans different types of **physical agents**. They argue that humans, animals, and robots all face conceptually similar challenges in physical reasoning, even if their bodies differ ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). For example, a person, a robotic arm, and a self-driving car all need to interpret sensory inputs, anticipate the results of their actions, and respect physics – just in different contexts. The ontology (shown as a table in the paper) crosses the key reasoning skills with various agent embodiments (natural agents like humans/animals vs. robotic systems like arms, humanoid robots, or vehicles) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). This highlights common sense reasoning as a general requirement for any physical entity interacting with the world. In effect, the authors are attempting to generalize common sense beyond a single domain: an AI's physical common sense should transfer across embodiments.

How is this theory implemented? **Cosmos-Reason1** is the model the authors develop as a step toward physically grounded AI reasoning (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). It's a large language model extended to handle **visual inputs (videos)** and to produce step-by-step reasoning. The model architecture is a decoder-only multimodal transformer: it uses a vision encoder to process video frames, projects those into the same embedding space as text tokens, and then feeds the sequence into an LLM that generates text outputs (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). During inference, Cosmos-Reason1 can observe a video and then **generate a chain-of-thought reasoning trace** in natural language, concluding with an **"embodied decision"** – for example, a description of the next action that should be taken in the scene (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) (Cosmos-Reason1: From Physical Common Sense To

Embodied Reasoning). This design ties perception to action via reasoning: the model sees what's happening, explains it to itself (the CoT process), and suggests an appropriate action, all in text form. Importantly, the output is not just a dry answer but often an explanation plus an action recommendation, reflecting the model's common-sense understanding of the situation. By working in the domain of video (rather than static images or text only), the model deals with **dynamic, time-based phenomena** – exactly the kind of real-world complexity where physical common sense is required. In summary, the authors connect their theory to the physical world by building an AI that learns from videos, reasons about space and time, and issues action guidance, thereby operationalizing common sense in an embodied context.

Benchmarks for Common Sense and Embodied Reasoning

To evaluate their theory and models, the authors created dedicated **benchmarks** that align with the defined ontologies. They note that existing AI benchmarks often fail to test true physical understanding (models might do well on standard datasets while still lacking grasp of real-world dynamics) (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). Thus, they constructed new test sets to specifically probe physical common sense knowledge and embodied decision-making.

For **physical common sense reasoning**, the team curated a question-answering dataset based on short internet videos ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). They began with over 5,700 candidate questions (covering a mix of binary yes/no and multiple-choice) about what is happening or what is possible in various video clips, then distilled this to a balanced set of **604 questions** on 426 distinct video scenarios ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). Each question is labeled according to the ontology categories from earlier (Space, Time, or Fundamental Physics) to ensure coverage of all aspects of common sense ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). For instance, a question might ask whether an object will remain hidden (testing object permanence, a Fundamental Physics concept) or what event caused a later event (testing temporal causality in the Time category). The final benchmark contains a mix of 336 binary questions and 268 multiple-choice questions, with a roughly 13/49/38% split across Space, Time, and Physics categories respectively ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). Figure 7 of the paper shows the distribution of question types, confirming that **a large portion targets temporal reasoning (which the authors identified as especially challenging for current models)** ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). **Notably, the questions are designed to require reasoning – the answer usually cannot be found by just looking at a single frame or reading subtitles, but by interpreting the physical interactions in the video over time** ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). The model must use common sense (e.g. understanding gravity or tool use) to arrive at the correct answer. The authors only measure whether the final answer is correct, and they "leave quantitatively assessing the quality of the thinking trace for future work" ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) – meaning the focus is on outcome accuracy, not yet on evaluating how "reasonable" the model's explanation was.

For **embodied reasoning**, the benchmark centers on action and planning questions. The authors constrain this evaluation to three key tasks that reflect an agent's interaction with the environment ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning):

1. **Task-completion verification:** determining if a given task or subtask in the video was successfully completed. (For example, "Did the person finish pouring the water into the glass?")
2. **Action affordance:** judging whether a specific action is possible or sensible in the situation. (For example, "Can the robot pick up the box in its current state?")
3. **Next plausible action prediction:** choosing the most likely or appropriate next step toward a goal. (For example, given a cooking video and current progress, "What should the cook do next?")

All embodied reasoning questions are presented as multiple-choice (often yes/no for the first two types, or a choice of possible next actions for the third) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). To build a comprehensive test, the authors pulled data from several sources that cover different embodiments and contexts. They took samples from **RoboVQA** (a robotic visual question answering dataset), **RoboFail** (a set of robot execution failure videos), **BridgeData** (robotic manipulation tasks with a gripper), **AgiBot** (agent-instruction videos), **HoloAssist** (augmented reality assistant scenarios), and an internal **AV (autonomous vehicle)** dataset ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). From each, they selected around 100 clips and formulated one question per clip focused on the above three reasoning tasks. They also took care to **standardize the question format and action granularity** across these sources ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). This means they phrased questions in a uniform way (so that the model must rely on the video, not on any dataset-specific wording quirks) and they ensured the possible answer choices are comparable in specificity. For example, when asking about the next action, all options are at a similar level of detail – avoiding a mix of very high-level and very low-level descriptions which could confuse the model ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). They even manually refined some multiple-choice options to remove ambiguities and ensure that solving the question truly requires understanding the video ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). The end result is an **embodied reasoning benchmark** totaling about **610 questions** (roughly 100 from each source, slightly more from RoboVQA) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). These cover a wide array of physical contexts: household tasks with human actors, robot-arm manipulation trials, human-robot interactions, instructional videos, and driving scenarios. By evaluating models on this benchmark, the authors can measure how well an AI generalizes its embodied common sense across different domains – from kitchen tasks to factory robots to self-driving cars.

Benchmark Results: The paper reports model performance on these new benchmarks for Cosmos-Reason1 and compares it to several existing models (including GPT-4's vision-enabled

variant, among others). As expected, prior models struggle with many of these physically grounded questions. **The authors point out that even very large multimodal models often barely do better than random chance on certain intuitive physics challenges like identifying the arrow of time (telling if a video is playing forwards or backwards) or tracking objects that disappear from view** (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). For instance, GPT-4 and a proprietary OpenAI model could solve spatial puzzles a bit better than chance, but **failed on temporal order and object permanence tests** (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). This suggests that traditional benchmarks haven't been adequately exercising these aspects of common sense (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). By contrast, the Cosmos-Reason1 models, after targeted training, achieve much stronger results. The Cosmos-Reason1-56B model (56 billion parameters) answered a majority of the benchmark questions correctly, outscoring the baseline models by a significant margin across both the common sense and embodied reasoning tasks (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). The authors highlight more than 10% accuracy improvement from their **Physical AI supervised fine-tuning**, and an additional ~8% boost from **Physical AI reinforcement learning**, as evidence that their approach yields tangible gains (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). Perhaps more interestingly, Cosmos-Reason1 demonstrates the ability to grasp concepts like **irreversibility of time and persistent object identity** which stump other AI systems. After training, it understands the "arrow of time" in videos (e.g. it can tell if a video of eggs unscrambling is physically implausible unless played in reverse) and maintains awareness of objects that temporarily vanish from sight, showing a learned sense of object permanence (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). These were explicit goals of the authors' curriculum, and the benchmark results confirm progress on them. In short, the benchmarks provided a way to **quantify common sense**, and Cosmos-Reason1's performance suggests that the authors' theory-driven training led to a measurable improvement in physical understanding.

Strengths of the Approach

- **Structured Theory of Common Sense:** A major strength of the paper is how it articulates what common sense means for physical AI. By breaking down common sense into ontologies of Space, Time, and Physics (with well-defined sub-capabilities), the authors provide a clear target for researchers and models ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). This is a novel contribution – instead of treating "common sense" as a vague catch-all, they enumerate concrete aspects of reality an AI should understand (from simple spatial relations to complex causal chains). Such a framework is valuable for guiding development and evaluation of AI systems going forward.
- **Integration of Embodied Reasoning:** The paper doesn't stop at theoretical definitions; it ties common sense to **embodied action**. By emphasizing perception and interaction (the System 1/System 2 mix, and the need for an agent to plan in a real environment), the approach acknowledges that true common sense is active. This combined perspective is innovative. Many past works looked at common sense purely as knowledge (often testing it with text questions), whereas here the authors insist that using knowledge in real-world

tasks is part of common sense ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). This holistic view (knowledge + action) aligns well with how humans employ common sense and is likely necessary for AI intended to operate in complex settings.

- **Comprehensive Benchmarking:** The creation of new benchmarks aligned with the theory is a strong point. The authors identified gaps in existing evaluations (e.g. models could do well on benchmarks yet lack basic physics intuition (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning)) and responded by crafting tests for those very gaps. The benchmarks cover a wide range of scenarios and task types, which means they can reveal an AI's weaknesses in understanding physical reality. This contribution is not just for their model – they have effectively provided the community with a diagnostic tool for physical common sense. Future researchers can use these benchmarks to measure progress, making the work impactful beyond the specific models introduced.
- **Demonstrated Performance Gains:** On the practical side, Cosmos-Reason1's results show clear improvements over prior systems on the targeted common sense tasks. The authors report double-digit percentage boosts in accuracy after fine-tuning on physical reasoning data, and further gains with reinforcement learning (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). In areas like intuitive physics, where GPT-4 and others fell short, Cosmos-Reason1 performs significantly better (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). For example, existing models struggled to do better than random chance on determining if events were temporally reversed, whereas Cosmos-Reason1 learned to reliably detect the correct temporal order (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). Such empirical gains lend credence to their approach – it's not just a philosophy of common sense, but a working implementation that advances the state of the art in embodied AI reasoning.
- **Open-Source and Reproducibility:** The authors indicate they will release their code and pretrained models under an open license (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). This openness is a strength because it allows others to build on their "physical common sense" theory and verify the results. Given the complexity of the system (ontologies, multi-stage training, etc.), providing the community with resources to replicate or adapt Cosmos-Reason1 is crucial for making a lasting impact. It also sets a positive example in a field where not all large models or datasets are available for scrutiny.

Limitations and Criticisms

- **Scope of "Common Sense":** While titled as a step toward a Physical Theory of Common Sense, the work deliberately restricts itself to **physical** commonsense knowledge and embodied tasks. It does not tackle social commonsense, language pragmatics, or other non-physical domains of common sense that are also important in AI. This narrow scope is sensible for focusing the research, but it means the resulting theory and model address only a subset of what humans consider "common sense." For instance, understanding that

others have beliefs and intentions, or that lying is possible (social common sense), is outside the purview here. The authors acknowledge that even within physical common sense, they haven't implemented learning from real-time interactions yet ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning) – arguably one of the most "common sense" things an agent can do is to learn from trial and error. So, the theory in its current form is incomplete: it's a foundation that will need to be extended to cover more aspects of general common sense.

- **Manual Curation and Ontology Design:** The approach relies heavily on human-designed structures (the ontologies) and human-curated data. The ontology of physical common sense, while comprehensive, is somewhat subjective – one could debate if the chosen categories truly span all of physical common sense or if they overlap. The need to pre-define 16 subcategories could be seen as a knowledge engineering approach in an era where many AI successes come from letting data speak for itself. Likewise, the datasets for fine-tuning and benchmarking were assembled and filtered through significant manual effort (thousands of questions written and vetted) ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). This raises a concern about **scalability and bias**: will the AI generalize beyond the concepts the designers thought of? If an unseen physical phenomenon comes up that doesn't fit neatly into those categories, the model might be confused. In short, the system's "common sense" is still bounded by the foresight of its creators. A truly robust common sense would ideally emerge more autonomously or cover an unlimited range of scenarios – something not yet achieved.

- **Simulation vs. Real Embodiment:** Despite focusing on embodiment, Cosmos-Reason1 operates in a **simulated manner**. It watches videos and answers questions; it does not physically intervene in the world. This is a pragmatic choice (training a large language model to output text is far easier than training a robot to act in the real world), but it means some challenges of physical agency are postponed. For example, the model doesn't deal with sensorimotor control, real-world noise in actuation, or the ethical and safety implications of taking physical actions. Its "embodied decisions" are still just text descriptions of actions. Thus, the **transfer from text output to real robot behavior** is an open question. The model might say "open the door gently," but executing that in a particular robot with actual force control is another level of problem. The paper's title hints at a physical theory, and indeed the conceptual groundwork is laid, but the implementation stops short of a full physical agent. Future work would need to connect the dots from Cosmos-Reason1's outputs to actual robotic controllers, and it's unclear how difficult that step might be.

- **Evaluation of Reasoning Quality:** A subtle issue is that, by the authors' own admission, they did not **evaluate the chain-of-thought reasoning traces** in a quantitative way ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). The model produces these multi-step explanations internally (which is great for interpretability), but success was measured solely by final answer accuracy. This means it's possible the model sometimes arrives at correct answers via flawed logic or lucky guesses – scenarios that the current evaluation wouldn't catch. For common sense, the process of reasoning is as important as the result, since one motivation is to trust the AI's

judgments. The authors note this as future work, and it stands as a weakness that currently the "common sense" can't be fully trusted just by looking at answer accuracy. A related point is that reinforcement learning was applied to reward correct answers, which could risk encouraging the model to produce convincing-sounding reasoning that leads to the right answer, rather than strictly truthful reasoning. In other words, there is a concern of potential **reasoning hallucination** that isn't penalized as long as the answer is right. Addressing this will require new techniques to evaluate and enforce consistency in the chain-of-thought itself.

- **Complexity and Resource Intensity:** The solution presented – two large LLMs (8B and 56B), a custom architecture with vision integration, multi-phase training (including a full RL pipeline) – is highly complex and resource-intensive. Training such models required aligning vision and language on millions of examples, then fine-tuning on specialized datasets, then running reinforcement learning with human-designed rewards. This approach might be out of reach for many research groups without significant compute resources. It also means the results combine many factors, and it can be hard to disentangle which pieces of their approach are truly necessary for common sense. For instance, did the hierarchical ontology mainly help with evaluation, or did it also improve training by structuring the data? Would a model half the size with the same data do nearly as well, or is scale critical? The paper doesn't fully isolate these variables (which is understandable, as their goal was to push for a high-performing system). The **practical impact** is that while Cosmos-Reason1 is a step forward, implementing one's own "common sense model" will still be a daunting endeavor following this recipe. In time, perhaps these techniques will distill down to simpler components, but currently the barrier to entry is non-trivial.
- **Generality and Future Adaptability:** Finally, a critical question is how well this physical theory of common sense will extend to new domains or evolving definitions of common sense. The benchmarks introduced are thorough, but any fixed benchmark can become gamed or overfitted. The real world always has surprises (think of the COVID-19 pandemic introducing new "common sense" about distancing, or new technologies changing how we interact physically). The ontology may need to evolve – e.g., if robots gain new sensors or abilities, what's "common sense" for them could expand. The authors have taken a significant step by formalizing and testing many concepts, and their model can in principle be further trained, but maintaining common sense will likely be an ongoing effort. The paper does not directly address how the model could update its knowledge post-deployment (since learning from interaction was left out). Thus, one weakness of the current approach is that it yields a static common sense model, one that might still make odd mistakes outside its training distribution and has no built-in mechanism to rapidly correct or learn from those mistakes on the fly.

In summary, Toward a Physical Theory of Common Sense presents a thoughtful framework for grounding AI reasoning in the physical world. The authors' key insight is to treat common sense not as an amorphous collection of facts, but as a set of capabilities that can be explicitly taught, evaluated, and incrementally improved. By connecting these capabilities to embodiment, they tackle the long-standing gap between abstract AI reasoning and real-world operation. The Cosmos-Reason1 model and its benchmarks illustrate both the promise and the challenge of this

endeavor: it's possible to significantly improve an AI's grasp of physical reality (the successes with intuitive physics are encouraging (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning)), yet true human-level common sense remains a distant goal. The work's strengths lie in its clear structure and demonstrated advancements, while its weaknesses highlight that we are still in the early days of understanding how to make AI learn and generalize common sense as effortlessly as humans do. Overall, this paper lays a solid foundation and is a substantial step toward AI that can **"perceive, understand, and perform complex actions in the physical world"** as the authors set out to achieve (Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning). The coming years will tell how this physical theory of common sense evolves and how it might integrate with other aspects of intelligence to yield AI agents that truly think and act with common sense in our rich, unpredictable world.

Sources: The summary above is based on **Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning** (arXiv:2503.15558) and associated content, including direct excerpts and figures from the paper ([2503.15558] Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning).