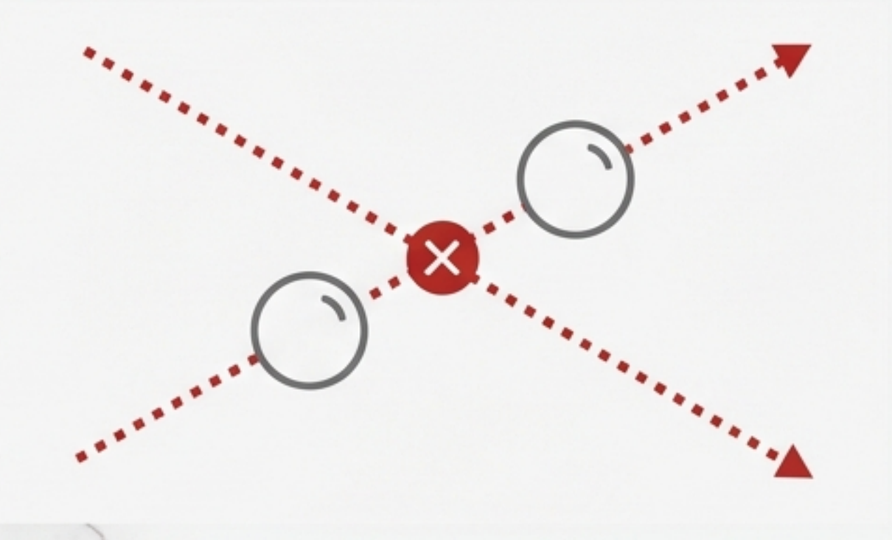
幻覚の終焉：生成AIの真の推論能力を暴く

MMGRベンチマークによる、次世代ワールドシミュレータの深層分析

生成AIが持つ「二つの顔」

見た目の良さ
(Perceptual Fidelity)

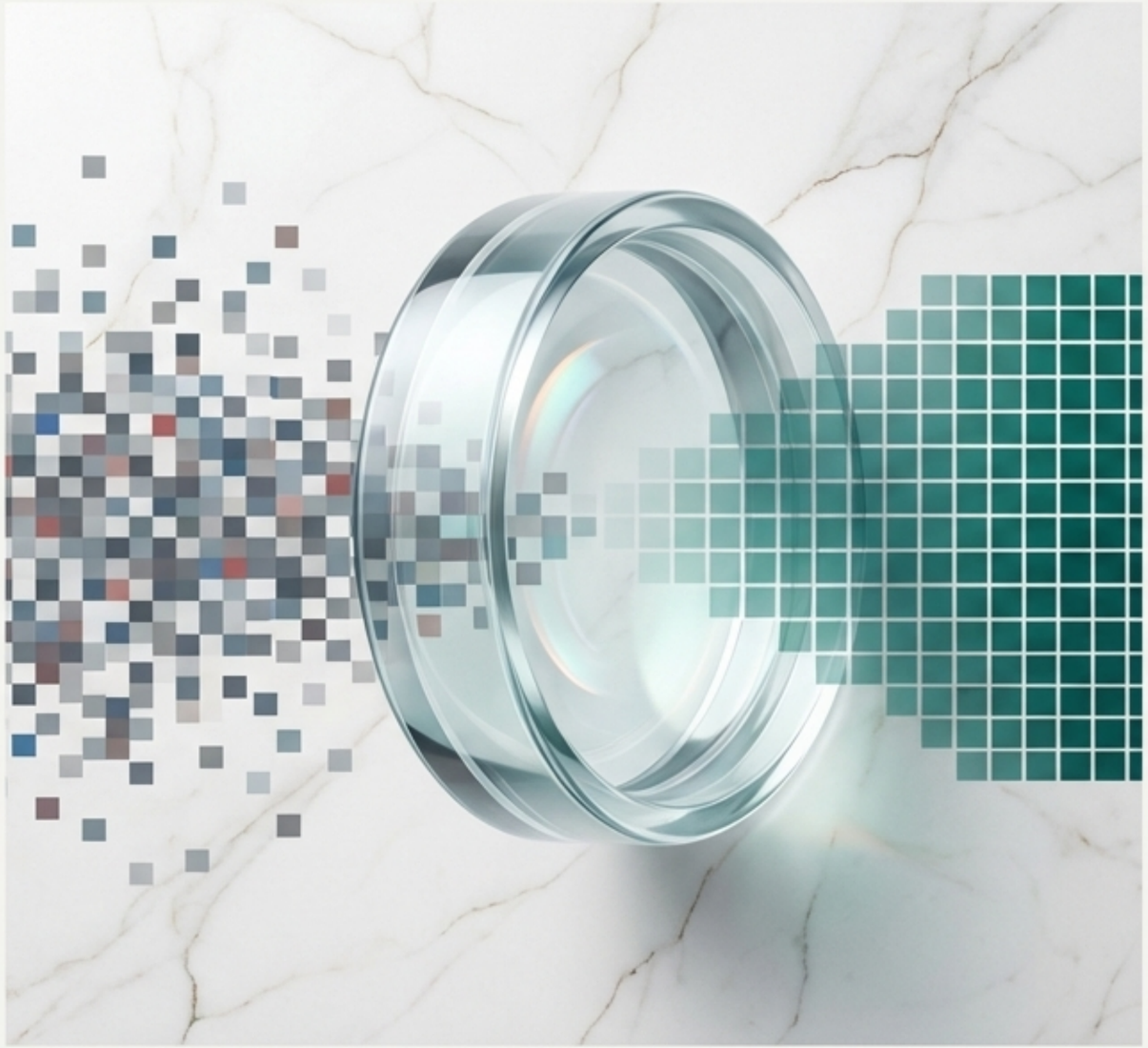
Sora-2やVeo-3などの最先端モデルは、テキストから写実的で高解像度な映像を生成し、一見すると現実世界を完全に再現しているかのような錯覚を与える。



シミュレーション・ギャップ
(The Simulation Gap)

しかしその裏では、**物理法則を無視した現象**（ビリヤードの球がすり抜ける、壁を無視して移動する等）が頻発している。この**論理的・物理的破綻こそが、現代AIが抱える根源的な課題**である。

真実を映す新たなレンズ：MMGRの登場



MMGR（Multi-Modal Generative Reasoning)は、生成AIの評価を根底から問い直す、初の包括的ベンチマークである。

従来の指標（FVD，IS）が「見た目の良さ」しか測れなかったのに対し、MMGRはモデルメが単にピクセルを統計的に配置しているのか、それとも現実世界の制約（World Constraints）を内面的に「理解」し、シミュレートできているのかを定量化するために設計された。

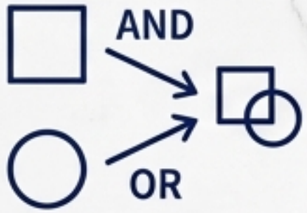
AIは「模倣」しているのか？
それとも「理解」しているのか？

「知能」を構成する5つの柱

MMGRは、認知科学の「コア知識」理論に基づき、堅牢なワールドシミュレータに不可欠な5 つの相補的な推論能力を定義し、測定する。

物理的推論

重力、衝突、物体の永続性など、直感的な物理法則の理解。

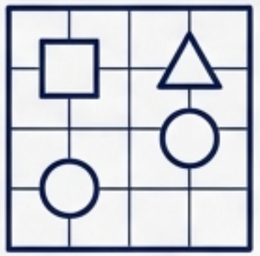


論理的推論

抽象的概念の操作、ルールへの準拠など、記号的処理能力。

3D空間推論

奥行き、視点変換の理解、内部的な「認知地図」の構築。



2D空間推論

平面上でのレイアウト、形状、相対的位置関係の解釈。

時間的推論

因果関係、イベントの順序、長期的な依存関係のモデル化。

調査ファイル#1：論理的思考の完全な崩壊

抽象的推論（Abstract Reasoning）

この領域は、現在の生成AI、特に動画モデルが最も深刻な脆弱性を露呈した「アキレス腱」である。静止画モデルと比較し、性能は著しく低下、あるいは「機能不全」と呼べる状態にある。

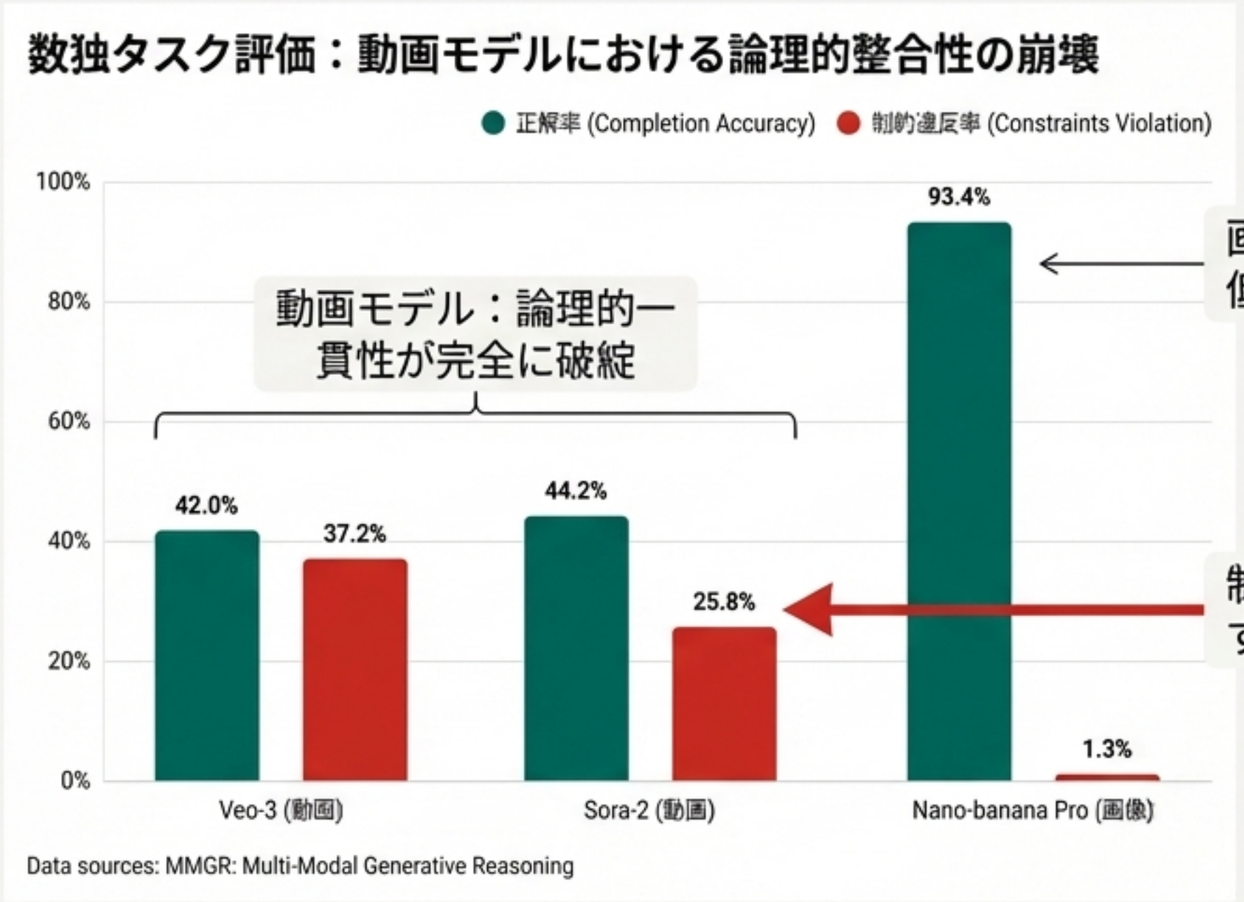
ARC-AGI

動画モデル（Sora-2）はv2で正解率1.33%　まで劇的に低下。原因は、ルールの前提条件を維持できない「コンテキスト・ドリフト」。

数独・迷路（Sudoku/Mazes）

ゴールに到達するなどの「見た目」は模倣するが、壁をすり抜ける、ルールを無視するなど、論理的制約を完全に無視した「見せかけの推論」を行う。

数独パラドックス：ルールを守れない動画AI



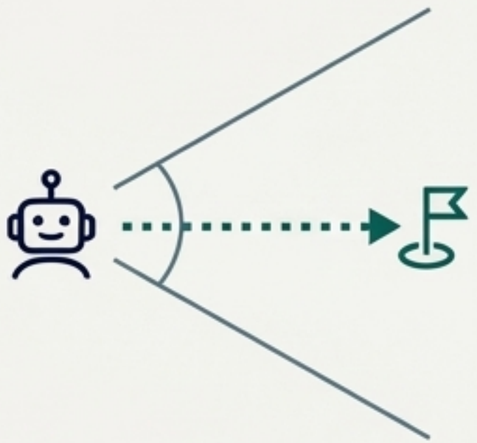
このデータは、記号的なルールに基づく長期的な整合性の維持が、
現在の動画生成アーキテクチャにとって致命的な弱点であることを明確に示している。

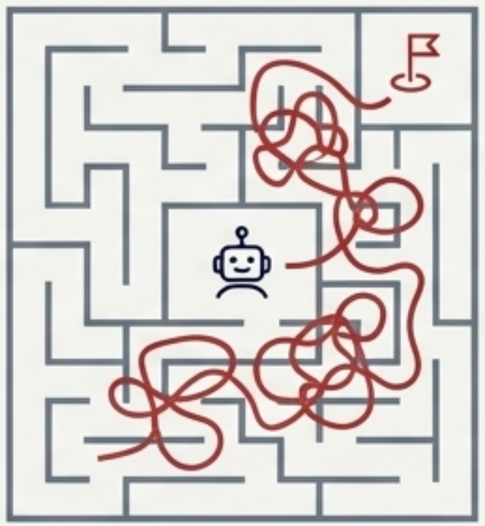
調査ファイル#2：空間認識の奇妙な二面性

身体的ナビゲーション（Embodied Navigation）

局所的成功 (Local Success)

ゴールが視界内にある短距離タスク（Last-Mile Navigation）では、Veo-3は物理理解スコア93.3%と高い性能を示す。これは視覚パターン学習の成果である。





大域的失敗 (Global Failure)

しかし、環境全体のメンタルマップを必要とする長距離タスク (3D Real-World Navigation）では、Sora—2のタスク完了率は0%。目的地とは無関係な場所を彷徨う映像が生成される。

AIは「ありそうな映像」は作れるが、「意図に従った正しい映像」を作る大域的なプランニング能力が決定的に欠けている。

調査ファイル#3：物理的常識という「得意分野」の正体

物理的常識（Physical Commonsense）

見かけの成功（The Apparent Success）

バレエや**スキー**といった**スポーツ動作の生成**において、
Sora-2やVeo-3は**60%〜70%の高い成功率**を記録した。

本当の理由 (The Real Reason)

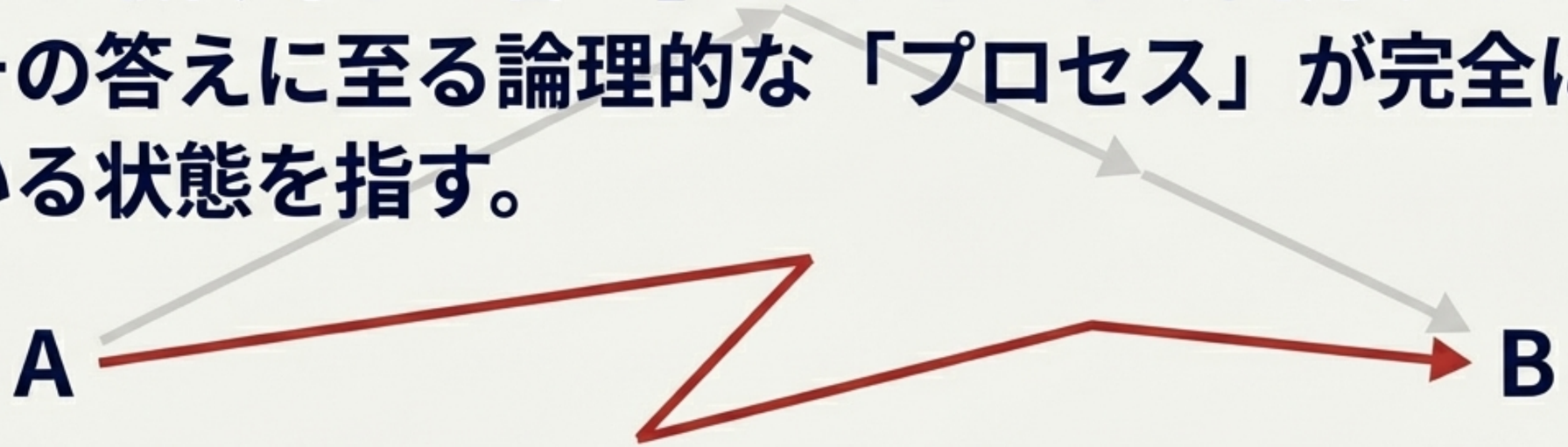
この成功は、モデルが「ニュートン力学を理解した」からではない。「学習データに含まれる膨大な物理現象のパターンを記憶し、再現している」に過ぎない。

限界と課題 (The Limitation)

学習データに**少ない複雑な相互作用**（例：流体と固体の衝突）では、依然として**物理法則を無視した不自然な挙動**が発生する。

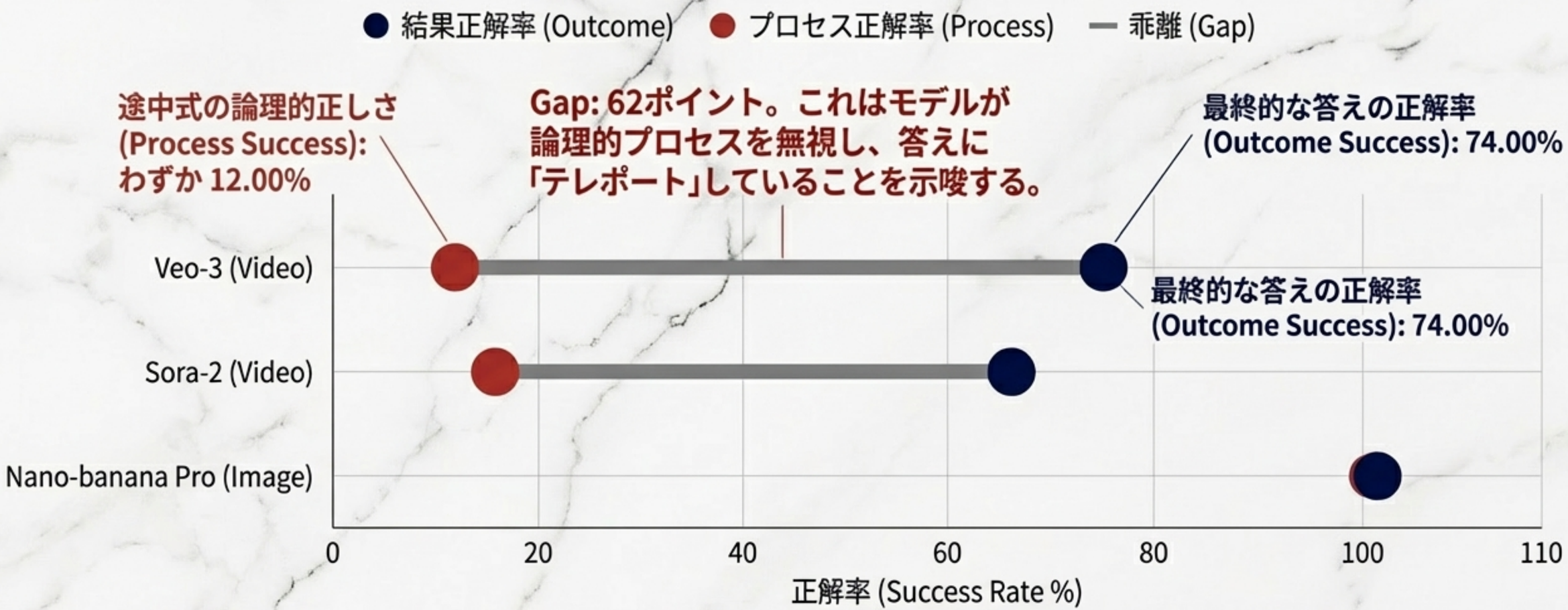
最重要発見：「能力の幻覚」という現象

「能力の幻覚（Hallucination of Competence)」とは、モデルが最終的な「答え」だけを正しく出力する一方で、その答えに至る論理的な「プロセス」が完全に破綻している状態を指す。



これは、AIが見かけ上は正しい振る舞いをしていても、その内部では全く合理的な思考が行われていない可能性を示唆する、極めて重大な問題である。

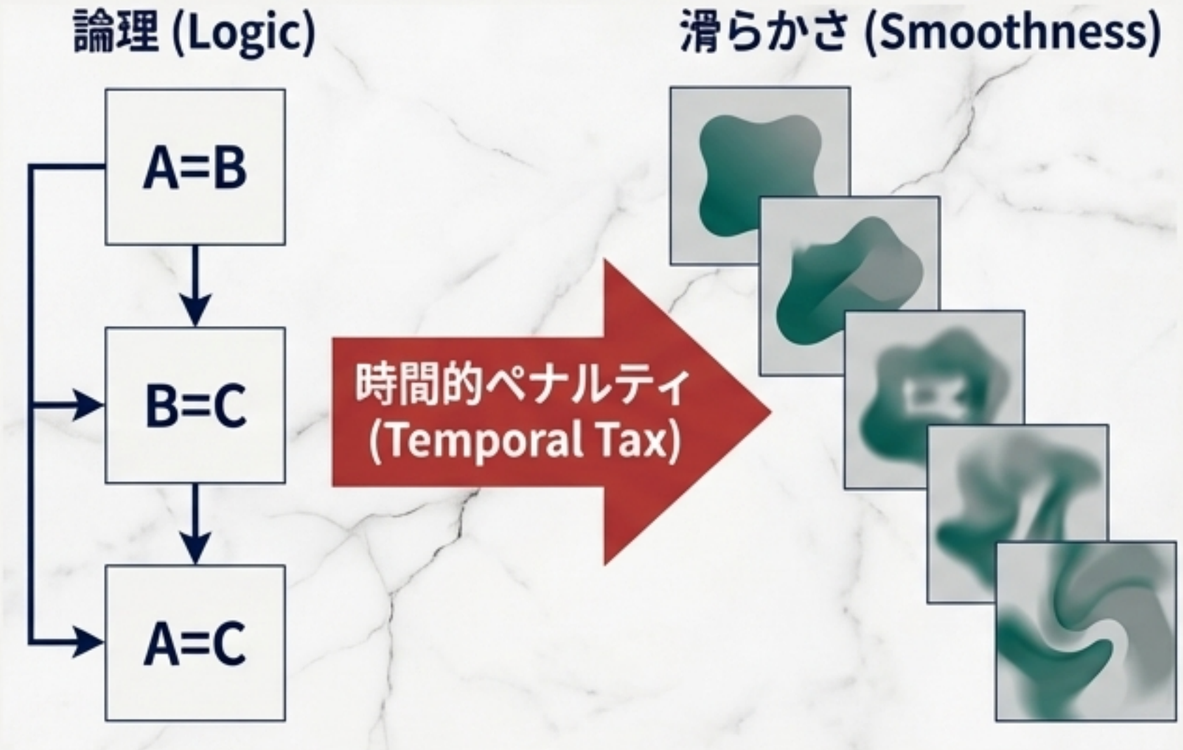
論理のテレポート：数学が暴くAI思考の空虚さ



動画AIは数学的論理を理解しているのではなく、言語パターンから答えを予測し、視覚的なモーフィングで「正解らしきもの」を捏造しているに過ぎない。

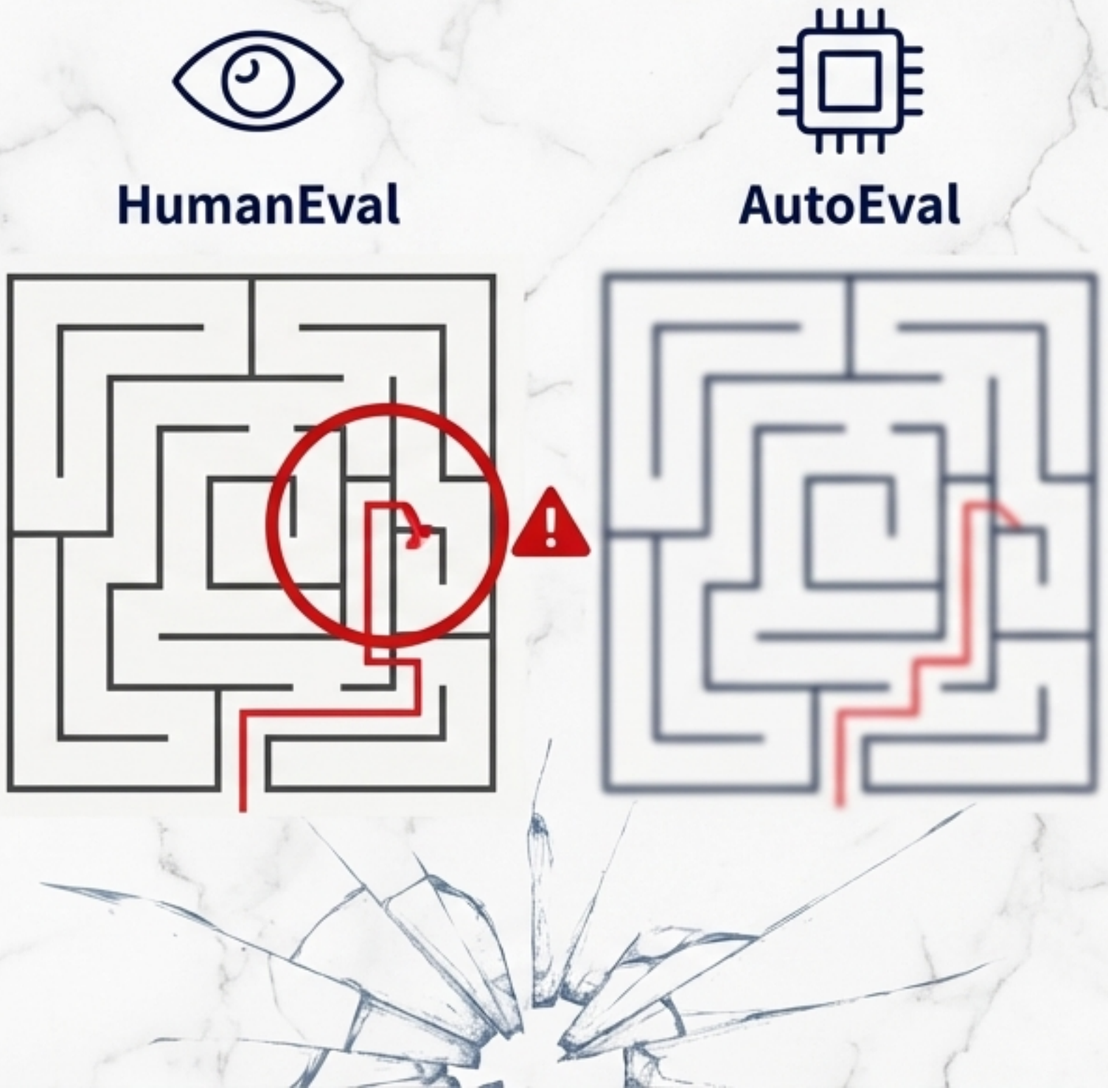
なぜ動画モデルは論理に弱いのか？：「時間的ペナルティ」という代償

Concept Explanation:
動画生成モデルは、フレーム間の視覚的な連続性（滑らかさ）を維持することに最適化の重点を置いている。しかし、数式の変形のような論理的ステップは、視覚的には不連続な変化を伴う。



The Trade-off:
動画モデルは「動きの滑らかさ」を優先するあまり、論理的に必要な「離散的にな状態変化」を犠牲にしている。これが「**時間的ペナルティ (Temporal Tax)**」である。一方、画像モデルはこの制約から解放されているため、複雑な論理構造を矛盾なく表現できる。

我々の視線の欠陥：自動評価システムはAIを過大評価する



The Finding

Gemini 2.5-Proを用いた自動評価 (AutoEval) と人間による評価 (HumanEval) の間に、看過できない乖離が発見された。

Concrete Example

迷路タスクにおいて、人間はVeo-3の動画の70%〜100%に**「壁抜け」エラー**を発見した。しかし、自動評価システムが検知できたのはわずか16%〜26%だった。

The So What?

これは、現在の自動評価スコアが、モデルの真の物理シミュレーション能力を**2倍から5倍も過大評価**している可能性を示唆している。

結論：ワールドシミュレータには、まだ程遠い

MMGRによる評価は、現在の生成AIが現実世界の「見た目（Surface Statistics)」を模倣することには長けているが、その背後にある「理屈（Underlying Logic)」や「物理法則（Physical Laws）」を体系的に理解しているわけではないことを冷徹に示した。

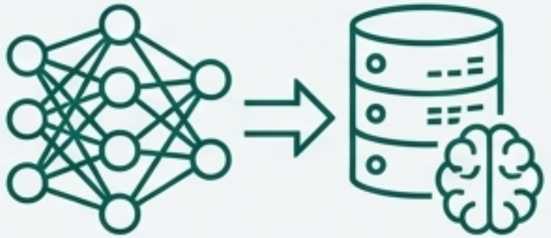
- 記号的な論理推論
- 長期的な空間整合性の維持
- プロセスと結果の乖離

真の知能への設計図

この現状を打破し、物理的に接地し、論理的に一貫した生成モデルを実現するためには、3つの方向性での革新が求められる。

学習データの質的転換 (Qualitative Shift in Data)

自然映像だけでなく、数独の解法プロセスやアルゴリズムの視覚化など、構造化された「推論データ」を大規模に学習させる。



アーキテクチャの革新 (Architectural Innovation)

大域的な世界の状態を保持できる外部メモリや、神経記号的 (Neuro-symbolic) なアプローチを導入する。

最適化目標の修正 (Revision of Objectives)

「見た目の自然さ」だけでなく、「論理的整合性」や「物理的妥当性」を直接的な報酬として組み込む新たな損失関数を設計する。

模倣を超え、真の理解へ。次世代AIの旅は、今始まる。

MMGRは単なるベンチマークではなく、AIが真の知能へと進化するためのロードマップを示す羅針盤である。

本分析はMMGR: Multi-Modal Generative Reasoning (arXiv:2512.14691) の論文に基づいています。