MMGR (Multi-Modal Generative Reasoning): 生成AIにおける「推論」と「物理的整合性」の包括的評価に関する深層分析レポート

Gemini 3 pro

1. 序論：知覚的模倣から認知的シミュレーションへの転換

人工知能研究、特に生成モデルの領域において、近年達成された技術的進歩は目覚ましいものがある。Text-to-Videoモデルの出現は、静止画の枠を超え、時間的連続性を伴う動的な視覚世界を合成する能力を人類にもたらした。Sora-2、Veo-3、Wan-2.2といった最先端モデルは、テキストプロンプ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の物理法則を完全に再現しているかのような錯覚を与える。しかし、この「見た目の良さ（Perceptual Fidelity）」の背後には、依然として埋めがたい「シミュレーション・ギャップ」が存在していることが、最新の研究によって明らかになりつつある[1]。

従来、生成モデルの評価はFréchet Video Distance (FVD) やInception Score (IS) といった指標に依存してきた。これらは生成された映像の分布が実際の映像分布とどれだけ類似しているか、あるいはテキストと意味的に整合しているかを測るものであり、映像内部の論理的整合性や物理的妥当性を問うものではない。その結果、モデルは「ビリヤードの球が互いをすり抜ける」「ナビゲーションエージェントが壁を無視して移動する」といった、現実世界ではあり得ない現象（幻覚：Hallucination）を平然と生成してしまう[1]。

本レポートで詳解する「MMGR (Multi-Modal Generative Reasoning)」は、この課題に対する根源的な問い直しである。MMGRは、モデルが単にピクセルを統計的に配置しているのか、それとも現実世界を支配する物理的、論理的、空間的な制約（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、初の包括的なベンチマークスイートである[1]。本分析では、MMGRの設計思想、各ドメインにおける詳細な実験結果、そしてそこから導き出される現代AIの認知的能力の限界と可能性について、徹底的な深掘りを行う。

2. 理論的枠組み：5つのコア推論能力と3つの評価ドメイン

MMGRの設計は、認知科学における「コア知識（Core Knowledge）」の理論に深く根ざしている。人間が乳児期から発達させる世界認識の基盤と同様に、堅牢なワールドシミュレータには、以下の5つの相補的な推論能力が不可欠であると定義されている[1]。

第一に「物理的推論（Physical Reasoning）」である。これは重力、摩擦、衝突、物体の永続性といった直感的な物理法則の理解を指す。第二に「論理的推論（Logical Reasoning）」であり、抽象的な概

念の操作、ルールへの準拠、条件分岐（もしAならばB）といった記号的処理能力、いわゆる「システム2」的な思考プロセスに相当する。第三および第四は空間認識に関わる能力であり、MMGRでは明確に区別されている。「3D空間推論（3D Spatial Reasoning）」は、奥行き、トポロジー、視点変換に伴う幾何学的整合性の理解であり、内部的な「認知地図」の構築を必要とする。一方、「2D空間推論（2D Spatial Reasoning）」は、投影された平面上でのレイアウト、形状認識、相対的な位置関係の解釈に関わる。最後に「時間的推論（Temporal Reasoning）」は、因果関係、イベントの順序、長期的な依存関係をモデル化し、過去の状態から未来の状態を一貫して生成する能力である [1]。

MMGRはこれらの能力を複合的に評価するため、相互に補完し合う3つのドメインを設定している。それぞれのドメインは特定の能力群に焦点を当てつつ、モデルの総合的な推論力を多角的に炙り出す構造となっている。まず「抽象的推論（Abstract Reasoning）」ドメインでは、迷路、数独、ARC-AGIといったタスクを通じ、物理世界から独立した純粋な論理的・2D空間的推論能力を測定する。次に「身体的ナビゲーション（Embodied Navigation）」ドメインでは、エージェント視点での環境理解を問い、物理的制約（壁の不透過性など）と3D空間認識、そして時間的な計画能力の統合を評価する。最後に「物理的常識（Physical Commonsense）」ドメインでは、日常的な物理現象やスポーツ動作の生成を通じ、直感的な物理法則の理解度を検証する [1]。

3. 抽象的推論 (Abstract Reasoning)：アキレス腱の露呈

MMGRの評価結果において、現在の生成AIモデルが最も深刻な脆弱性を露呈したのは、この抽象的推論の領域である。特に、動画生成モデルにおいては、静止画モデルと比較して著しい性能低下、あるいは「機能不全」とも呼べる状態が確認された [1]。

3.1 ARC-AGIにおけるコンテキスト・ドリフトと記憶の欠落

ARC-AGI（Abstraction and Reasoning Corpus）は、少数の入出力例から潜在的な変換ルールを推論し、未知のテストケースに適用する能力を測るタスクであり、AIの「流動性知能」を測る試金石とされている。このタスクにおいて、最先端の動画モデルであるSora-2でさえ、比較的容易なバージョン1（v1）での正解率は20.18%にとどまり、より複雑なパターンを含むバージョン2（v2）では1.33%へと劇的に低下した。これに対し、画像モデルであるNano-banana Proはv1で30.54%、v2でも30.36%と安定したスコアを維持しており、モダリティ間で明確な「推論格差」が存在することが判明した [1]。

動画モデルが失敗する主要なメカニズムとして、「コンテキスト・ドリフト（Context Drift）」が特定されている。ARC-AGIのようなタスクでは、問題の提示部分（例示画像）は不変のルールを示す「静的な真実」として保持されなければならない。しかし、動画モデルは時間経過とともに、この例示部分の色やパターンを勝手に変形させたり、グリッド構造を崩壊させたりする傾向が強い [1]。これは、現在の動画生成アーキテクチャが「時間的な変化（動き）」を生成することに過度に適応しており、論理的な前提条件としての「不変性」を維持する機構、すなわち長期的なメモリやグローバルな状態管理能力が欠如していることを示唆している。

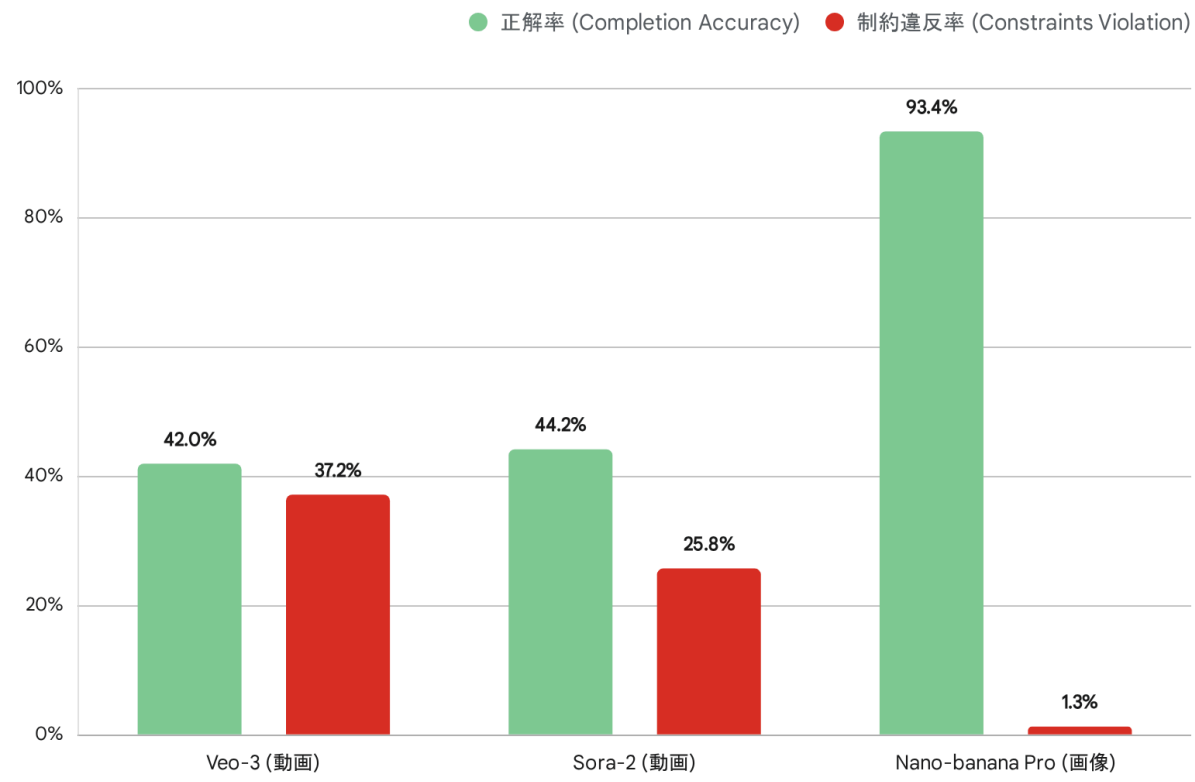
3.2 数独と迷路における「見せかけの推論」

数独や迷路といったタスクにおいても、動画モデルの挙動は興味深いパラドックスを示した。Veo-3などのモデルは、迷路のスタートからゴールへ向かうエージェントの動きを生成すること自体には成

功し、高い「目標達成率（Target Achievement）」を記録する場合がある。しかし、その過程を詳細に分析すると、エージェントが壁をすり抜けて移動する「Cross Wall」エラーが頻発していることが判明した。Veo-3の場合、Easyレベルの迷路でさえ、自動評価では壁抜け率が約25%と判定されたが、より厳密な人間による評価では70%〜100%のケースで壁抜けが発生していた [1]。これは、モデルが「迷路を解くアルゴリズム」を実行しているのではなく、学習データ内の「ゴールに到達した状態」という視覚的な結果パターンのみを模倣し、物理的な障壁という論理的制約を無視していることを意味する。

数独タスクにおいても同様の現象が見られる。動画モデルは、空欄に数字を書き込んだり、一度書いた数字を修正したりするような「推論のような振る舞い（Action Reflection）」を視覚的に生成する。しかし、その修正プロセスは論理的な整合性に基づいておらず、前のフレームで正しく書き込んだ数字を次のフレームで矛盾する数字に書き換えてしまうといったエラーが多発する。結果として、Veo-3の数独タスクにおける全体的な成功率は、人間評価において0%であった [1]。

数独タスク評価：動画モデルにおける論理的整合性の崩壊



画像生成モデル（Nano-banana Pro）は高い完了精度と低い制約違反率を示す一方、動画生成モデル（Veo-3, Sora-2）は制約違反率が極めて高く、論理的な一貫性を保てていないことがわかる。特にSora-2は制約違反が100%に近い。

Data sources: MMGR: Multi-Modal Generative Reasoning

このデータ（上図参照）は、画像モデル（Nano-banana Pro）が高い完了精度と極めて低い制約違反率を両立しているのに対し、動画モデル（Veo-3, Sora-2）は完了精度が低いだけでなく、制約違反率が異常に高いことを示している。特にSora-2に至っては、制約違反率が100%に達するケースも確認されており、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が浮き彫りになった [1]。

4. 身体的ナビゲーション (Embodied Navigation)：空間認識のパラドックスとモダリティの壁

身体的ナビゲーションドメインでは、エージェントが3次元環境内を移動する動画を生成させることで、3D空間の理解度、経路計画能力、および物理的制約の遵守状況を評価する。ここでもまた、モ

デルの挙動には興味深い「二面性」が観察された。

4.1 局所的成功と大域的失敗：視界の中と外

Veo-3などの最新モデルは、一人称視点の「ラストワンマイルナビゲーション（Panoramic View Last-Mile Navigation）」のように、ゴールが現在の視界に入っている短距離タスクにおいては、比較的高い性能を示した。具体的には、物理的理解スコア（Object Semantic Scoreなど）で93.3%という高い数値を記録している[1]。これは、学習データセットに含まれる膨大な一人称視点の移動映像（ゲームプレイ動画や実写映像）から、局所的な奥行き認識、オプティカルフロー、物体回避といった基本的なナビゲーションスキルを、視覚的なパターンとして効果的に学習できていることを示唆している。

しかし、この能力はあくまで「視界内」に限定されているようである。フロアをまたぐ移動や、見えない目的地への長距離移動など、環境全体のメンタルマップ（認知地図）を必要とするタスクにおいて、性能は急激に低下する。「3D実世界ナビゲーション（3D Real-World Navigation）」タスクでは、Sora-2のタスク完了率（Overall Success）は0%であった[1]。モデルは、もっともらしい廊下や部屋の映像を生成し続けることはできるが、指定された目的地に向かうという「操縦性（Steerability）」や「意図の維持」が欠如しており、結果として目的地とは無関係な場所を彷徨う映像が生成される。これは、生成AIが「ありそうな映像（Plausible）」を作る能力には長けているが、「特定の意図や地図に従った映像（Correct）」を作る能力、すなわち大域的なプランニング能力において決定的な欠陥を抱えていることを示している。

4.2 SLAGタスクと視点間の乖離 (Cross-View Alignment)

MMGRで導入された「SLAG（Simultaneous Localization and Generation）」タスクは、モデルに3D視点での移動映像と、それに対応する2D俯瞰マップ上の軌跡を同時に生成させるという、極めて高度な課題である。このタスクの結果は、モデルの空間認識の限界をさらに明確にした。

Nano-bananaのような画像モデルは、静的な状態での3Dビューと2Dマップの整合性は高いレベルで維持できる。しかし、動画モデルにおいては、時間の経過とともにエージェントの3D空間内での位置と、2Dマップ上に描画される軌跡との間にズレが生じ、最終的には全く異なる場所を示してしまう「乖離」現象が確認された。Veo-3の全体成功率は11.2%、Sora-2は12.9%にとどまり、Wan-2.2に至っては0.8%とほぼ破綻している[1]。

一方で、興味深い現象も観察された。SLAGのように、3D映像と2Dマップという異なるモダリティを同時に生成させるという制約（Cross-View Constraints）を課した場合、単独で3D映像のみを生成させた場合と比較して、Veo-3のシーン一貫性（Scene Consistency）スコアが向上したのである[1]。これは、マルチモーダルな出力要求がモデルに対して強い制約として働き、潜在的な空間推論能力を部分的に引き出した可能性を示唆している。しかし、テキストによる指示（「赤い部屋に行け」など）のみでナビゲーションを行う場合、モデルのパフォーマンスはさらに低下し、言語的な空間指示を視覚的な行動計画に接地（Grounding）させることの難しさが改めて確認された。

5. 物理的常識 (Physical Commonsense)：唯一の得意分野

とその背景

「抽象的推論」や「身体的ナビゲーション」における苦戦とは対照的に、「物理的常識」ドメインは、現在の生成モデルが最も高いスコアを記録した領域である。

5.1 学習データのバイアスと「記憶」による解決

バレエの回転動作、スキーの滑降、水泳のストローク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%程度の高い成功率を示した[1]。また、基本的な物理現象（物が落ちる、転がるなど）についても、比較的自然な映像を生成することができている。

この成功の主要因は、モデルが「ニュートン力学を理解した」からではなく、「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ている」ことにあると分析される。現在の動画生成モデルの学習データセットには、YouTubeや映画、ストックフッテージなど、人間や物体が物理法則に従って動く自然映像が大量に含まれている。モデルはこれらのデータから、人体関節の可動域や重力下での物体の放物線運動といった統計的な規則性を獲得しており、それを再現することで「物理的に正しい」ように見える映像を生成しているのである。

5.2 複雑な相互作用における限界

しかし、この「記憶に基づく物理」には限界がある。学習データに頻出するパターン（人が歩く、走る）は得意だが、データに含まれにくい、あるいは微細な条件の違いが結果を大きく左右するような複雑な相互作用においては、依然として失敗が見られる。例えば、流体と固体の相互作用（水しぶきの正確な挙動）や、複数の物体が複雑に衝突し合うシナリオ（積み木崩しなど）では、物理法則を無視した挙動や不自然な変形が発生することがある。特に、Wan-2.2はスポーツシナリオにおいても成功率が21.33%と低迷しており[1]、モデルの規模や学習データの質によって、この「物理的記憶」の精度にも大きな差があることがわかる。

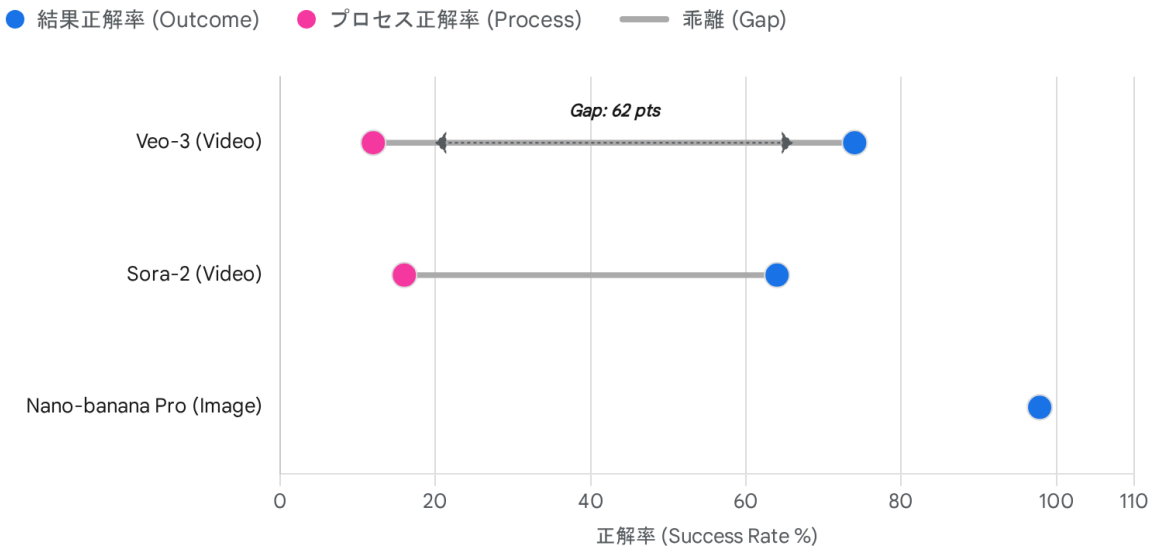
6. 生成AIにおける「能力の幻覚」：プロセスと結果の乖離

MMGRの評価結果から得られた最も衝撃的かつ重要な洞察の一つは、「能力の幻覚（Hallucination of Competence）」と呼ぶべき現象の発見である。これは、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。

6.1 数学問題における「結果オーライ」現象と論理のテレポート

Visual Mathタスクにおいて、動画モデルはこの現象を顕著に示した。GSM8K（小学校レベルの文章題）を用いた評価において、Veo-3は最終的な答えの正解率（Outcome Success）で74.00%という高い数値を記録した。しかし、その答えに至るまでの中間生成プロセス、つまり途中式や図解の変遷が論理的に正しいかを評価したプロセスの正解率（Process Success）は、わずか12.00%に過ぎなかった[1]。

「能力の幻覚」：数学タスクにおける結果とプロセスの乖離



Veo-3（動画モデル）は**結果正解率**と**プロセス正解率**の間に巨大な乖離（約62ポイント）があり、論理的プロセスを無視して答えを出していることが示唆される。一方、Nano-banana Pro（画像モデル）は両者が一致しており、着実な推論を行っている。

Data source: MMGR: Multi-Modal Generative Reasoning (arXiv:2512.14691)

上図が示す約62ポイントもの巨大なギャップは、モデルの内部処理の実態を雄弁に物語っている。モデルは問題文に含まれる言語パターンから「答えとなる数字」を直接予測することには成功している。しかし、その数字を導き出すための視覚的な推論ステップ（例えば、方程式の変形や図形の操作）を生成する際には、数字が魔法のように変形したり、文脈と無関係な計算式が突然出現したり、あるいは論理的な飛躍を経て最終的な正答に「テレポート」するような挙動を見せる。

これは、現在の動画生成AIが数学的論理を理解しているように見せかけて、実際には言語モデル的なパターンマッチングと、視覚的なモーフィング技術の組み合わせによって「正解らしきもの」を捏造しているに過ぎないことを示唆している。教育現場や科学的シミュレーションなど、結果だけでなくプロセスそのものの正確性が求められる分野において、このような「結果オーライ」のAIを利用することは重大なリスクを伴う。

6.2 時間的ペナルティ（Temporal Tax）

なぜ動画モデルは、画像モデルに比べてこれほどまでに論理的タスクで劣るのか。本レポートでは、その原因として「時間的ペナルティ（Temporal Tax）」という概念を提唱したい。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、チラつきのなさ）を維持することに計算リソースと最適化の重点を置かなければならない。しかし、論理的なステップ（例：数式の変形）は、視覚的には不連続な変

化を伴うことが多い。

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「離散的な状態変化」を犠牲にしている可能性がある[1]。これに対し、画像モデルは静止画一枚の生成に全リソースを集中でき、時間的整合性という制約から解放されているため、複雑な論理構造を一枚の画像内に矛盾なく配置することに成功している。動画モデルが「動き」を作ろうとして「論理」を犠牲にしているというこのトレードオフは、今後のアーキテクチャ設計における重要な課題となるだろう。

7. 評価手法の課題：自動評価と人間評価の乖離

MMGRの実験過程では、モデルの性能評価手法そのものについても、重大な発見があった。Gemini 2.5-Proを用いた自動評価（AutoEval）と、訓練された人間による評価（HumanEval）のスコアを比較したところ、特定のタスクにおいて看過できない乖離が見られたのである。

7.1 自動評価システムの「甘さ」と物理的盲点

自動評価システムは、特に「一時的な物理違反」を見逃す傾向が強いことが判明した。例えば、迷路タスクにおいて、人間はVeo-3が生成した動画の70%〜100%に「壁抜け（Cross Wall）」という物理的エラーを発見した。しかし、VLMベースの自動評価システムは、同じ動画群に対して16%〜26%しかエラーを検知できなかった[1]。

この原因は、現在のVLMが主に静止画の解析能力に基づいており、動画を評価する際にもフレームごとのサンプリングに依存しているためと考えられる。エージェントが高速で移動し、壁をすり抜ける瞬間がサンプリングの間に起きた場合、あるいはその変化が一瞬である場合、VLMはその物理的矛盾を捉えきれない時間的分解能の限界を露呈する。

7.2 評価の信頼性と評価可能性（**Evaluability**）

この結果は、現在の自動評価スコアが、実際のモデル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。特に、物理的な正しさが絶対条件となるタスクにおいては、依然として人間による検証（Human-in-the-loop）が不可欠である。

また、この問題は「評価可能性（Evaluability）」という新たな視点も提供する。Nano-bananaのような画像モデルは、移動の軌跡を静的な青い線として一枚の画像に描画するため、その経路が壁と交差しているかどうかを一目で検証することが容易である。一方、動画モデルは情報を時間の経過とともに分散して提示するため、評価者（人間であれAIであれ）にとって認知負荷が高く、エラーの見落としを誘発しやすい。モデルが自身の推論プロセスや意図を、検証可能な形式（例：Chain-of-Thoughtsの視覚化や、軌跡のオーバーレイ表示）で出力する能力を持たせることは、単なる性能向上だけでなく、AIの信頼性と評価の透明性を高めるためにも極めて重要である[1]。

8. 結論：ワールドシミュレータへの道程

MMGRによる包括的かつ深層的な評価は、現在の生成AIモデルが真の「ワールドシミュレータ」と呼ぶには時期尚早であることを冷徹に示した。モデルは現実世界の「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。特に、記号的な論理推論や長期的な空間整合性の維持においては、驚くほど脆弱であることが明らかになった。

8.1 今後の研究開発への提言

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

1. 学習データの質的転換: 現在のデータセットは自然映像に偏重しており、論理構造を学習するためのデータが圧倒的に不足している。数独の解法プロセスや、アルゴリズムの視覚化、物理実験の精密な記録など、構造化された「推論データ」を大規模に学習に取り入れる必要がある[1]。
2. アーキテクチャへの記憶と推論の統合: 現在のTransformerやDiffusionモデルは、局所的な相関関係の学習に最適化されている。大域的な世界の状態（Global World State）や不変のルールを保持・参照できる外部メモリ機構や、ニューラルネットワークと記号推論を融合させた神経記号的（Neuro-symbolic）なアプローチの導入が、コンテキスト・ドリフトを防ぐ鍵となるだろう[1]。
3. 最適化目標の修正: 「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」を直接的な報酬シグナルとして組み込む新たな損失関数（Loss Function）や強化学習フレームワークの設計が不可欠である。

MMGRは単なるベンチマークにとどまらず、生成AIが単なるコンテンツ生成ツールから、現実世界を理解し推論する真の知能へと進化するためのロードマップを提供する羅針盤である。この評価軸が示す課題を一つずつ克服していくプロセスこそが、次世代のAI研究の主戦場となるであろう。

引用文献

1. 2512.14691v2.pdf