

MMGR (Multi-Modal Generative Reasoning) 包括的技術解説

1. ベンチマーク設計思想 – なぜMMGRが必要か

従来の画像・動画生成モデル評価は見た目のリアリティに偏重しており、生成物の物理法則や論理的一貫性までは評価できていませんでした¹。例えば、AIが生成したビリヤードの動画でボール同士が幽霊のようにすり抜けたり、ナビゲーション動画でロボットが壁を瞬間移動で通り抜けても、Fréchet Video Distance (FVD) やCLIPスコアなど既存指標では高評価になり得ます²。FVDは生成動画の分布が実データに近い(知覚的類似度)を見る指標、CLIPベースの評価はテキスト記述との整合性を見る指標ですが、いずれも**因果関係の破綻や物理法則の無視**といった「推論の失敗」を検知できません¹。このような課題感から、「見た目が本物らしいだけでは不十分で、本当に世界の制約を理解しているか」を問う評価基準が求められました。

MMGR (Multi-Modal Generative Reasoning) は、このギャップを埋める初の包括的ベンチマークです³。モデル内に**現実世界の物理的・論理的・空間的な制約の理解**が内在しているかどうかを評価することを目的とし、以下の5つのコア推論能力を測定軸としています⁴：

1. **物理的推論** – 重力や衝突、物体の恒常性といった**直感的物理法則**の理解⁵。例えば物が消えずに持続する、一度投げたボールは落下する、といった因果を守れているか。
2. **論理的推論** – ルールに従った**記号的・抽象的な推論能力**⁵。例えば「もしAならばB」といった論理、数独パズルのような規則操作などの**シンボリック処理**に相当します。
3. **3D空間推論** – **三次元空間内の関係**やナビゲーション、トポロジーの理解⁶。カメラ視点が変わっても物体の位置関係を維持したり、マップを内部表現して経路を計画する能力です。
4. **2D空間推論** – **二次元平面上のレイアウト**や形状、相対位置の正確な解釈⁷。画像平面で物体同士の上下左右関係をきちんと守り、指定されたレイアウトを崩さずに配置できるかなど。
5. **時間的推論** – **因果関係やイベントの順序**といった時間的整合性のモデル化⁷。動画の前後関係が論理的に通っているか、長期的な依存関係(ストーリーの一貫性、フレーム間の状態保持)を維持できるかがポイントです。

以上のようにMMGRは、「**モデルが世界をどれだけ一貫してシミュレートできるか**」を測ることを設計理念としています。従来指標では見逃されていた**因果構造の破綻、物理法則違反、大局的な不整合**を露呈させることで、生成モデルの本当の賢さを評価しようとするものです¹²。

2. 評価構成 – 3つのドメインとタスク例

MMGRは上記5つの推論能力をバランスよくテストするため、**3つの評価ドメイン**にタスクを分類しています⁸。各ドメインは互いに補完的で、異なる種類の推論を要求します。

- ### **抽象的推論 (Abstract Reasoning)** – 物理世界から切り離された**論理パズル系**のタスク群です⁹。例として**迷路問題**、**数独**、**視覚的な数学パズル**などが含まれます。図形や記号のパターンを操作して解答を導くものであり、主に **論理的推論** と **2D空間推論** (および一部の時間的推論) を評価します⁹。具体的には、与えられた迷路画像でスタートからゴールへの経路を示す動画を生成させたり、不完全な数独盤面から解を完成させる一連の画像を生成させる、といった課題が出されます。中

でもARC-AGI (Abstraction and Reasoning Corpusの拡張版) ¹⁰ や数独は典型例で、モデルが純粋な記号ルールに従った推論を行えるかを問うものです。

- ### 身体的ナビゲーション (Embodied Navigation) – エージェント視点のナビゲーションに関するタスク群です ¹¹。カメラ視点の動画内でエージェント (例: 人やロボット) が建物内を移動し、目的地へ到達する過程を生成させる課題など、**3D空間推論・2D空間推論・時間的推論**および**物理的推論**が複合的に要求されます ¹¹。例えば「建物の入口から指定された部屋までの経路を示す動画を作れ」というタスクでは、モデルは内部に地図的な世界モデルを持ち、**長期的な空間計画**を映像で表現できるかが試されます ¹¹。このドメインには、実写に近い3D環境 (GibsonやHabitatのような実環境シミュレーション) でのナビゲーション、上空視点のマップでの経路描画、同時**自己位置推定とシーン生成** (SLAM的課題) など複数の設定が含まれ ¹² ¹³、空間把握と運動計画の一貫性が評価されます。
- ### 物理的常識 (Physical Commonsense) – 日常物理やスポーツなど物理法則に根ざしたタスク群です ¹⁴。例えば「ボールを投げてゴールに入れるシーン」や「コップの水を注いで満たすシーン」のように、**直観的な物理法則**に則った動画を生成できるかを評価します ¹⁴。重力に従って物体が落下する、液体が容器の形に沿って流れる、物と物が衝突すれば跳ね返る等、**物理的推論と時間的一貫性**が主に問われます ¹⁴。スポーツ映像の生成では、ボールの軌道や人間の動きが自然か、物と物との相互作用 (接触や摩擦) が現実的か、といった点が評価されます。また複数物体の**合成的な相互作用** (例えばドミノ倒しの連鎖など) も課題に含まれ、モデルの因果推論能力を検証します。

以上の3ドメインを合わせて、MMGR全体では**合計1,853サンプル**からなる「試験問題集」のような評価スイートになっています ¹⁵。各タスクごとに**細粒度の評価指標**が設定されており、単にそれっぽい出力を1フレーム生成するだけでなく**動画全体を通した整合性**や**問題の正解率**まで測られる点が特徴です ¹⁶。例えば迷路タスクでは「ゴールに到達したか」「壁をすり抜けていないか」「途中で迷路そのものを改変していないか」といった指標で**総合正解**が判定されるなど、部分的な成功ではなく**ホリスティックな正解**を要求します ¹⁶ ¹⁷。

3. 実験結果 – 現行モデルの性能と誤りパターン

MMGR論文では、最新の画像・動画生成モデル計8種ほどがこのベンチマークで評価されています ¹⁸。動画モデルでは**Veo-3** (Google/DeepMindのモデル)、**Sora-2** (OpenAIのモデル) など、画像モデルでは**Nano-Banana** (匿名化された高性能な拡散モデル) や**GPT-4o-Image**、**Qwen-Image** (それぞれGPT-4系・Alibaba Qwen系の画像生成) などが含まれます ¹⁸。これら最先端モデルでも、**推論が要求される場面で大きな性能ギャップ**が露わになりました。

- **物理常識ドメイン**では比較的健闘し、モデルはスポーツシーンなどで60%以上の成功率を示すケースもありました ¹⁹。例えば「ボール投げ」や「液体を注ぐ」といった映像生成は、訓練データ中に現実世界の物理現象の動画が豊富なため概ねそれらしく作られています ¹⁹。実際、MMGR全タスク中ではこの領域がもっとも高いスコアとなり、**直感的物理**についてはモデルが一定の知識を獲得していることが示唆されました。
- **抽象的推論ドメイン**では惨憺たる結果で、成功率は一桁%台に留まりました ¹⁹。特に**数独**や**ARC-AGI**パズルでは**10%未満 (ほぼランダムレベル)**の正解率しか得られず、ほとんど解けていません ²⁰。また**数学問題** (方程式を解く過程を動画で示すなど) では、画像モデルのNano-Banana Proが複雑な数学論理をかなり高精度でこなせた一方で、動画モデルのVeo-3は**Outcome (最終答の正否)**と**Process (過程の正当性)**が大きく乖離し、最終答だけ当てずっぽうに合って過程は滅茶苦茶という事例が頻発しました ²¹。例えば**中学算数テスト (GSM8K)** ではVeo-3が最終答を74%の高確率で当てる一方、途中計算が正しい割合はわずか12%に過ぎず、**見かけ上答えは合っているが証明になっていない**という状態でした ²¹。

- **ナビゲーション系タスクも大きな弱点**として露呈しました。長い経路を辿る課題では、途中で現在地を見失ったりゴールと無関係な場所にワープしたりといった不整合が頻発し、**長時間の空間計画**を維持するのが苦手です²²。例えば3Dナビゲーションでは、短い経路ではそこそこ正しく進めても、フロアを跨ぐような複雑経路では**全体成功率20%前後**にまで落ち込みました²³。モデルはローカルにはそれらしい映像を繋げますが、大局的にはゴールに辿り着けず整合性が崩れるのです²³。総じて、「**一貫した長尺のプランニング**」が今の生成モデルには困難であることが示されています。

具体的な誤り例も数多く報告されています。典型的なものを挙げます：

- **途中状態の破綻（矛盾）** - 数独の動画生成では、解答過程で盤面に最初なかった数字が突然現れたり、逆に与えられていたヒント数字を途中で書き換えてしまったりする例が観察されました²⁴。本来固定のはずの初期条件を勝手に改変して辻褃合わせをするなど、**論理一貫性の崩壊**が見られます。このように**問題の前提を維持できない**ことは、モデルがパズルのルールを内部で保持していない証拠です。
- **物理法則の違反** - 「コーヒーミルで豆を挽く」シーンを想像させると、豆が刃に触れた瞬間に**砕けず突然粉に変化してしまう**といった出力が得られました（まるでクロスフェードで豆が消えるような描写）²⁵。摩擦や質量保存といった基本法則が守られず、滑らかな映像効果でごまかされています。また、ビリヤードでは**球と球が衝突せずすり抜ける**例も多数ありました²。見た目はリアルなだけに、これらの**不自然な挙動**は「一見もっともらしいけど現実では有り得ない」違和感を生みます。
- **空間的一貫性の欠如** - ナビゲーションでは、モデルが**突然ワープ**するような動きをしがちです。例えばロボットが廊下を歩いていたかと思うと、次の瞬間目的の部屋の中に瞬間移動している、といった映像です²。ゴールに到達させるために**途中経路をすっ飛ばす**傾向が見られ、壁やドアを無視して移動するケースが頻発しました。このような**テレポーテーション**は長いシーケンスを整合させることの難しさを表しています。
- **ルール違反による「正解」** - 2D迷路タスクでは、**モデルが迷路そのものを書き換えて解いてしまう**例もあります²⁶。具体的には、ゴールへの経路が存在しない迷路で**壁を一部消去して抜け道を作る**、あるいは**ゴール位置を改変する**などの不正解動作です。画像生成型のモデルではこの傾向が顕著で、最終フレームではゴールに辿り着いたように見えても、プロセスを見ると**壁抜けや迷路改変**が行われており結果的に無効となる例が多発しました²⁷²⁶。実際、ある中難易度迷路では**画像モデルが82.5%の高確率でゴールに到達したものの、25%のステップで壁を通り抜けていたため有効解はわずか2.5%だった**という報告があります²⁸。一方、動画モデルはゴール到達率は低め（50%程度）でも**壁抜けは抑えて解こうとする**傾向がありました²⁸。これは動画モデルが「見た目の繋がり」を重視し**大幅な改変を避ける**一方、画像モデルは**各ステップ独立に最適化**するため無茶をしがちな違いと考えられます。

以上の結果から、「現在の生成モデルはうまくいっているように見えて、実は問題を解けていない」ケースが多々あることが分かりました。モデルによっては**最終結果だけ正しい答えを示し、途中の理屈はデタラメ**という挙動も散見されます²¹。これは後述する“**推論の錯覚**” (Illusion of Reasoning) とも関係する現象で、出力だけ見ると賢そうなのに内実は辻褃が合っていないという、AIのハリボテ的賢さを如実に示しています。

4. MMGRが映す生成AIの限界 – “推論の錯覚”と“時間的コスト”

MMGRの分析により、現在の生成AIモデルが抱える**根本的な限界**がいくつか浮き彫りになりました²⁹。主なポイントは次のとおりです。

- **データ偏りによる認知能力の穴:** 学習データセットの偏重がモデルの得意・不得意を生んでいます³⁰。すなわち「動画・画像の大量データには現実世界の物理現象が豊富だが、**記号的・論理的なデータが極端に少ない**」という不均衡です³¹。その結果、モデルは物理直感には身につけても**記号操作の経験が圧倒的に不足し**、数独やARCのような抽象課題ではほぼ白紙状態になります³⁰。実際、ARC-AGIや数独で10%未満という成績は、人間で言えば「**問題の意味すら分かっていない**」レベルであり²⁰、これはモデルの世界知識がいかに偏っているかを示しています。逆にスポーツ映像などは大量の訓練事例から**丸暗記に近い形で対応**できていると考えられます。このように**知識の濃淡**により特定の認知領域が穴になることが確認されました。
- **“推論の錯覚”と局所最適化:** モデルのアーキテクチャ上の限界として、**局所的なもっともらしさを優先しすぎて大局的な一貫性を犠牲にする傾向**があります³²。各フレームや部分的な画像では破綻がないように生成できても、長い時間軸や広い空間スケールで見ると辻褄が合わなくなるのはこのためです²³。例えばVeo-3はナビゲーション全体の指標で80%近い部分成功を収めても、**シーン全体で整合が取れた成功（総合成功率）は20%程度**しか達成できませんでした²³。これは、モデル内部で状態を保持・更新するメモリが弱く、**その場その場の映像を尤もらしくすることに注力**してしまうためです³²。その結果、**連続するフレーム間で論理矛盾が蓄積**し、例えば先述のように「いつの間にかパズルの条件が変わっている」「中盤で計算過程がおかしくなる」といった**推論破綻**に至ります。モデルは**自分が何を表現していたか忘れてしまう**のです。これがいわゆる**“Illusion of Reasoning”**（推論の錯覚）であり、AIモデルがあたかも考えているように見えても実際には整合的な推論はしておらず、**見かけの体裁を整えているだけ**であることを意味します²¹。実験では、Veo-3が**62%もの問題で「途中過程は減茶苦茶だが最後だけ正解を出す**」という振る舞いを示し³³、まさに**見せかけの解答**をしていたことが報告されています。
- **“時間的コスト”（Temporal Tax）問題:** 特に動画生成では、**時間的一貫性を保つこと自体が推論能力に対するコスト**になっていることが指摘されました³⁴。これを著者らは**“Temporal Tax”**と呼んでいます。つまり、フレーム間の見た目の繋がりを維持しようとする制約が重荷となり、論理的整合性とのトレードオフが発生しているのです³⁵。実験では、難易度の高い数学問題で動画モデルの成績が**画像モデルの1/4以下（4~6倍の性能差）**にまで落ち込むケースがありました³⁴。これは「**動的に解答過程を描く**」という負担が、推論そのものの正確さを著しく損なっていることを示唆します。要するに、**長い時間軸にわたって正しく考え続けることにモデルは大きなペナルティを支払っている**のです。この“Temporal Tax”は、時間方向に長いコンテキスト保持が難しい現行アーキテクチャの限界と言えます。現在の動画生成モデルは連続フレームを**逐次的に予測**するため、フレーム間整合性を守りつつ論理まで通すのは二重の負担になり、結果として**「滑らかならデータラメ」**な映像を生む傾向があるわけです。
- **評価指標と訓練目標のミスマッチ:** モデルが何を最適化して訓練されているかも重要な限界点です³⁶。現在の生成モデルは主にピクセルの再現誤差や敵対的訓練（GAN）で**見た目のリアルさ**を追求する目的関数を持っています³⁷。そのため、**論理的正しさや因果的一貫性は直接報奨されず**、極端な話「破綻があっても人間が気づきにくければOK」と学習してしまいます³⁶。実際、モデルは迷路問題で「**ちゃんと解く**」より「**それっぽく見える絵を描く**」ことを優先しており、壁を破ってでもゴールに向かう映像を作っていました³⁸。このように**最適化目標のギャップ**がある限り、モデルは**見かけ倒しの解答**を生成し続けるでしょう³⁶。MMGRが提起した課題は、評価だけでなく**学習目標の再考**が必要だという点にもあります。「正しく推論できたら報奨を与える」ような仕組み（ルール遵守や因果関係の維持にペナルティを科す訓練など）の検討が不可欠です³⁹。

以上のように、MMGRによって**生成AIの限界領域**が明確化されました。それは裏を返せば、今後この分野で克服すべき技術課題とも言えます。単に高解像度でフォトリアルなだけではなく、**世界のルールを理解した上で想像力を働かせるAI**へと進化させるために、どのような方向に研究を進めるべきか示唆が得られたのです⁴⁰。

5. 今後の研究・設計への示唆 – 世界状態の明示、メモリ、推論とレンダリングの分離

MMGRの著者らは、現状の課題を踏まえて**今後のマルチモーダル生成AIに求められる能力・アプローチ**をいくつか提言しています。

- **明示的な世界状態の表現:** モデル内部に**環境の状態を表す構造**を持たせることです。現在のモデルはピクセル単位で次を予測するため、物体やエージェントの状態を**暗黙的にしか保持していません**。そこで、シーン中のオブジェクトの配置や関係を明示的に記憶・更新できる**ワールドモデル**を組み込むことが提案されています⁴¹。例えば、物体ごとの位置・速度・属性をテーブルで管理し、それに基づいて画像を生成するような仕組みです。世界状態を持てば、時間が経過しても「**何がどこにあるか**」を忘れず、突然物が消えたり移動したりする矛盾を防げます。また新たな状態推論（物理法則の適用など）にもその情報を使えるため、**物理エンジン的な推論**を内部で行うことも可能になるでしょう。
- **外部メモリの活用:** 長い生成過程で**コンテキストを保持**するため、モデルに**外部メモリ**やリカレントな状態保持機構を持たせる必要性も指摘されています⁴¹。現在の巨大モデルは基本的に巨大な一枚の変換器で動作しており、内部メモリは自己注意メカニズム任せです。しかしこれでは長い系列では注意が行き届かなくなります。そこで、画像や動画の生成途中に**要約された状態をストア**し、後から参照できるような外部メモリを導入する研究が考えられます⁴¹。たとえば「迷路の構造」「これまで訪れた場所」「現在保持しているパズルの制約」などを都度書き込み読み出しできれば、長期一貫性が飛躍的に向上する可能性があります。このアイデアは言語モデル分野でも検討が進んでおり、同様にマルチモーダル分野でも**メモリ強化**が鍵となるでしょう。
- **推論モジュールとレンダリングモジュールの分離:** **論理推論のプロセス**と**画像・動画描画を切り離すアーキテクチャ**への転換も提案されています⁴²。現在は一つのモデルが「次のピクセル」を決める中で暗に推論もしている状態ですが、これを**二段階プロセス**にするイメージです⁴²。まず**推論エンジン**が内部でシンボリックな解答手順や物理シミュレーション結果を導き出し、それを元に**レンダリングエンジン**が映像化する、という役割分担です。こうすることで、推論エンジンは視覚的自然さに引きずられず**論理的な一貫性**だけを追求でき、レンダリングエンジンは推論結果に**忠実に見た目を作り込む**ことに専念できます⁴³。MMGR論文でも「推論状態と可視化をデカップリングせよ」と述べられており、具体的には**推論状態を表す中間表現**（テキストや構造データ）を介して画像を生成するような新アーキテクチャが展望されています⁴²。
- **学習目標の見直し:** 前述のように、モデルが**論理・物理の正しさ**を評価軸に学習していないことが問題でした。そのため、**報酬デザイン**を見直す提案もなされています⁴⁴。具体的には、ルール違反や因果矛盾にペナルティを与える**補助目的関数**を導入し、モデルがそれを避けるよう最適化する方向です⁴⁴。例えば生成結果をチェックする別AI（もしくはルールベースの判定器）を用意し、「迷路の壁を貫通していないか」「矛盾したフレーム遷移になっていないか」を判定して減点する、といった仕組みが考えられます。これは強化学習（Reinforcement Learning）や**人間のフィードバック**の活用などによって実装可能です³⁹。さらには**ニューラルシンボリック手法**（推論をソフトなシンボル操作で行わせる）との組み合わせも示唆されています⁴⁵。要は、モデルに「**正しく解け**」という動機付けを与えることで、見かけ倒しではない本当の推論能力を引き出そうという方向です。

これらの提言は、既にいくつかの研究で兆しが見えています。例えばVoxPoserという手法では、大規模言語モデル（LLM）と視覚モデルを組み合わせて**ロボット操作のための3D価値マップ**を生成し、それを使ってモーションプランニングするというアプローチを示しました⁴⁶。LLMが与えた指示から**制約やアフォーダンス**を推論し、3次元空間上のマップ（ここに行け、ここは避けよという空間的価値の地図）を構築、それを元にロボットの動きを決定します⁴⁶。これはまさに**外部の推論モジュール**（LLM）が**内部状態**（価値マップ）を明示的に作り、**レンダリング**（ロボットの動作）はプランナー任せ、という構造になっています。MMGRが指摘したような**世界状態の明示や推論と実行の分離**の有効性を裏付ける例と言えるでしょう。

総じて、MMGRは現行モデルの弱点を暴くだけでなく、**次世代マルチモーダルAIへの設計指針**を与えるものとなりました⁴⁰。「**世界を内部に持ったAI**」、すなわち物理法則や論理をネイティブに理解・シミュレートできるAIへのロードマップが示されたのです。今後はこれに沿って、メモリ拡張やモジュール分離、学習目標の工夫などが進み、より賢く信頼できる生成AIが登場することが期待されています⁴⁰。

6. 関連ベンチマーク・周辺技術との比較

MMGRと類似あるいは補完的な立場にあるベンチマークや技術として、**ARC-AGI**、**Gibson**（および類似の**Embodied AI環境**）、**VoxPoser**、**TouchStone**などが挙げられます。それぞれの概要とMMGRとの関係を整理します。

- **ARC-AGI** (Abstraction and Reasoning Corpus – Artificial General Intelligence) : フランソワ・シヨレによるARCは、人間の直感に近い抽象推論問題集として知られます⁴⁷。ARC-AGIはその発展版で、AIの汎用的な問題解決能力を測る難関ベンチマークです。**色と形のパターン操作**など純粋に記号的な課題で構成され、人間でも発想力が要求されます。MMGRの抽象推論ドメインにはこのARC-AGIから派生したタスクが含まれており⁴⁸⁴⁹、モデルがどれだけ**シンボリックな一般常識**を持っているかをテストしています。ARC-AGI自体は**出力もシンボル（画像）**で評価するものでしたが、MMGRでは**その過程を動画で示させる**点が異なります。結果として、ARC-AGI単体では測れなかった**解答プロセスの整合性**まで暴き出すことができました。現行モデルはARC-AGI問題の最終答すらほとんど当てられず（<10%）⁵⁰、この分野の弱さを再確認することに。ARC-AGIとMMGRは**目的は共通（推論力評価）**ですが、MMGRが**生成過程の評価**まで踏み込んだ点で革新と言えます。両者は相補的であり、ARC-AGIで培われた難問をMMGRが新たな角度から解かせることで、より深くモデルの思考力を検証できています。
- **Gibson環境**（および**Embodied AIナビゲーション・シミュレータ**） : Gibsonは実世界の3Dスキャン環境を用いたロボットナビゲーション用シミュレータで、Embodied AI（身体性を持つAI）の代表的ベンチマーク環境です。エージェント（ロボット）が実環境に近い仮想空間を移動し、ゴールまでナビゲートできるかを評価します。従来、この種のベンチマーク（例: **Habitat**やGibson, Matterport3D等）は**認識と経路計画**が中心で、AIはセンサー入力を基に動作を決めます⁵¹。評価指標もナビゲーション成功率や距離など**行動面**でした。一方、MMGRのナビゲーションドメインは「**映像を生成できるか**」にフォーカスしており、モデル内部で地図構築や経路計画ができているかを**視覚的アウトプット**から評価します⁵¹。言わば、Gibsonが**ロボット工学的性能**（正しくゴールに行けるか）を測るのに対し、MMGRは**創発的な環境理解**（頭の中でシミュレーションできているか）を測る違いがあります。実際、Gibson等ではシミュレータが物理法則を保証しますが、MMGRではモデル自身が物理法則を守る必要があり、その分難易度が高い設定です。MMGRナビゲーションで露呈した壁抜けやレポート問題は、Gibsonのような従来ベンチマークでは発覚しなかったタイプのエラーです²。したがって、**Gibson系ベンチマークとMMGRは協動的**であり、前者が**ロボット制御の実績**を評価するなら、後者は**脳内シミュレータとしてのAI**を評価する、と位置付けられます。将来的にはGibsonで鍛えたエージェントの知識をMMGRの評価で確認する、といった活用も考えられます。
- **VoxPoser** (ボックスポザー) : 前述したように、VoxPoserはStanford大学らによる**ロボット操作のためのLLM+VLMハイブリッド手法**です⁴⁶。LLM（大規模言語モデル）が命令文から論理推論を行

い、VLM（視覚言語モデル）と連携して3次元の**価値マップ**を構築、それをプランナーが読み取ってロボットアームの操作軌道を決めます⁴⁶。この方法論は、MMGRが指摘した「**推論とレンダリングの分離**」「**外部メモリ活用**」に通じるものです。つまり、VoxPoserでは**コード（プラン）**という中間表現を介しており、LLMが思考しVLMが環境を把握し、その結果を別モジュールが実行するという**モジュール分割**が実現されています⁴⁶。MMGRで浮かび上がった問題に対する**一つの解決策の方向性**を示すものと言えます。もっとも、VoxPoser自体はベンチマークではなく**技術提案**ですが、その成功は**明示的推論+世界モデル**の有効性を裏付けています。MMGR的な評価で求められる、「物理的に辻褃の合う動きをゼロから組み立てる」能力に対し、VoxPoserは**言語モデルの知識を活用して即興でプランを立てる**という別アプローチで応えており、今後このような**マルチモーダル推論と行動のブリッジング**が重要になることを示唆しています。

- **TouchStone**（タッチストーン）：2023年に提案されたTouchStoneは、**ビジョンと言語のマルチモーダルモデルを評価する新手法**です⁵²。特徴は**強力なLLM（GPT-4など）を審査員として用いる点**で、画像やマルチ画像入力に対するモデルの応答をテキスト化し、その内容の正しさや質をLLMが採点します⁵²。例えば画像についての質問応答を行わせ、モデルの回答とGPT-4が生成した模範解答を比較し、GPT-4にどれだけ合致するかでスコア付けするといった具合です⁵²。TouchStoneは**視覚的理解・物語創作・複数画像推論**など5次元の能力評価データセットからなり、人手による詳細アノテーションとLLMジャッジの組み合わせで**効率的かつ高精度な評価**を実現しています⁵³⁵⁴。MMGRとの違いは、MMGRが**明確な正解があるタスク**に対し**専用の自動指標や人間評価**で正誤を見るのに対し、TouchStoneは**オープンエンドなタスク**（自由応答）にLLMを用いた**相対評価**をする点です。TouchStoneは評価の自動化という意味で画期的ですが、LLM自体が完璧ではないため**評価者のバイアスや言語モデルの限界**に影響される可能性があります。一方MMGRは、迷路ならゴール到達や違反の有無といった**客観指標**が定義されており、評価軸が明瞭です⁵⁵。両者は**アプローチが補完的**で、TouchStoneのように**LLMをメタ評価に使う手法**は、人手では難しい創造的タスクの評価に有用でしょう。実際、MMGRが対象としない**対話的・創造的要素**（例えば物語生成の整合性など）はTouchStoneがカバーしています⁵²⁵⁶。逆にMMGRは**視覚的な厳密さ**に踏み込んだ評価をするので、物理や論理の厳格なチェックには適しています。したがって、将来的にはMMGRでハードな客観問題を解かせ、TouchStone的手法で自由応答の質も測る、といった組み合わせでモデルを多角評価することが望ましいでしょう。

このように、MMGRは他のベンチマーク・技術と**競合するものではなく補完関係**にあります。ARC-AGIから問題設定を受け継ぎ、Gibsonのようなシミュレータでは見えない生成時の不整合を暴き、VoxPoser的アプローチに解決策のヒントを得つつ、TouchStoneのような評価自動化とも両立し得る枠組みです。MMGRは「**理解から生成へ**」という新たな評価パラダイム⁵¹を示しました。他ベンチマークと連携しつつ発展することで、**真に論理と物理に通じたマルチモーダルAIの実現**に寄与していくと期待されます。

参考文献・情報源: MMGR 論文¹²⁵⁷⁹¹⁴¹⁹¹⁷³⁸²¹³³³⁴²³⁴¹⁵⁸；関連技術・ベンチマーク⁴⁶⁵²ほか。

¹²¹⁰²¹²³³³³⁴³⁵³⁷³⁹⁴¹⁴²⁴³⁴⁴⁴⁵⁴⁸⁴⁹⁵⁰⁵¹⁵⁸ 2512.14691v2.pdf

file:///file_000000007eac7209a9a0608f6b57b954

³⁴⁵⁶⁷⁸⁹¹¹¹⁴¹⁷¹⁹²⁰²²²⁶²⁷²⁸²⁹³⁰³¹³²³⁶³⁸⁴⁰⁵⁷ 【論文】 【AI】 マルチモーダル生成AIの「推論能力」を測る新ベンチマークMMGR | MASAKING

https://note.com/r7038xx/n/n662c41323f6a

¹²¹³ 2512.14691v2.pdf

file:///file_00000000e3e47209987de0b20270a88e

15 25 **Why AI Can't Do Physics: The MMGR Paper Exposes the Flaws in ...**

<https://ninza7.medium.com/why-ai-cant-do-physics-the-mmgr-paper-exposes-the-flaws-in-multi-modal-generative-reasoning-5c9632a1b44f>

16 18 55 **[2512.14691] MMGR: Multi-Modal Generative Reasoning**

<https://arxiv.org/abs/2512.14691>

24 **The Reality Gap: Why Your Video AI is a World-Class Artist but a Failing Physicist | by ArXiv In-depth Analysis | Dec, 2025 | GoPenAI**

<https://blog.gopenai.com/the-reality-gap-why-your-video-ai-is-a-world-class-artist-but-a-failing-physicist-02b809f8ac3e?gi=7cdd6e486008>

46 **VoxPoser**

<https://voxposer.github.io/>

47 **What is ARC-AGI? - ARC Prize**

<https://arcprize.org/arc-agi>

52 53 54 56 **GitHub - OFA-Sys/TouchStone: Touchstone: Evaluating Vision-Language Models by Language Models**

<https://github.com/OFA-Sys/TouchStone>