SMART: 次トークン予測によるスケーラブルなマルチエージェントリアルタイムシミュレーション

Wei Wu*
清華大学センスタイム研究 wuwei@senseauto.com

Xiaoxin Feng*
SenseTime Research
fengxiaoxin@senseauto.com

Ziyan Gao*
SenseTime Research
gaoziyan@senseauto.com

Yuheng Kan
SenseTime Research
kanyuheng@senseauto.com

Abstract

データ駆動型自律走行運動生成タスクは、データセットサイズの制限やデータセット間のドメインギャップの影響を頻繁に受け、実世界のシナリオでの広範な適用を妨げている。この問題に対処するために、我々はSMARTを導入する。SMARTは、ベクトル化された地図とエージェントの軌跡データを離散的なシーケンストークンにモデル化する、新しい自律走行モーション生成パラダイムである。これらのトークンは、デコーダのみの変換器アーキテクチャで処理され、空間-時間系列にまたがる次のトークン予測タスクのために学習される。このGPTスタイルの手法により、モデルは実際の運転シナリオにおける運動分布を学習することができる。SMARTは、生成的Sim Agentsチャレンジにおいて、ほとんどのメトリクスで最先端の性能を達成し、Waymo Open Motion Dataset (WOMD)のリーダーボードで1位を獲得し、顕著な推論速度を実証した。さらに、SMARTは自律走行領域における生成モデルを表し、ゼロショット汎化能力を示す: 訓練にNuPlanデータセットのみを、検証にWOMDのみを使用した場合、SMARTはSim Agentsチャレンジで0.72の競争力のあるスコアを達成した。最後に、複数のデータセットから10億以上のモーショントークンを収集し、モデルのスケーラビリティを検証した。これらの結果は、SMARTが当初、スケーラビリティとゼロショット汎化という2つの重要な特性をエミュレートし、大規模なリアルタイムシミュレーションアプリケーションのニーズを予備的に満たしていることを示唆している。我々は、自律走行分野における運動生成のためのモデルの探索を促進するために、全てのコードを公開した。ソースコードはhttps://github.com/rainmaker22/SMARTにある。

1 Introduction

自律走行の文脈では、ベクトル化された地図と車両の軌跡データを活用することで、運動計画[17, 6, 19, 18, 7]、運動予測[47, 11, 38]、Sim Agents[14]などの様々な運動生成タスクが容易になる。先行研究[9, 5, 27]では、運転シーンを表現するためにエンコーダネットワークを、マルチモーダルモーションを生成するためにデコーダネットワークを主に採用している。これらの生成された運動は、ガウス[4]またはラプラス[53]混合損失関数を用いて、連続的な軌跡分布に直接回帰される。このフレームワークは、回帰精度を優先する予測タスクでは強力な性能を示すが、プランニング[3]やシムエージェント[26]のような運転行動の安全性と合理性を重視するモーション生成タスクでは、しばしば性能が劣る。

*Equal contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

この性能不足の主な理由は以下の通りである: 第一に、このフレームワークは、異なるエージェントの動き間の将来の相互作用を表現していないため、シーンレベルの予測に一貫性がない。第二に、このモデルはデコーダで複数の意図クエリを初期化することでマルチモーダルなモーションを生成するが、これは一般的にGPUメモリによって制限され、その結果、モーションモダリティの数は一定となる。その結果、生成されたモダリティが将来の行動の多様性を十分に表現しているかどうかは不明である。第三に、これらのモデルは異なるデータセット間で汎化するのに苦労しており、新しい都市環境や地図での学習には新しいデータ収集が必要である。

自己回帰大規模言語モデル(LLM)[12, 42]の登場は、人工知能の新時代を切り開いた。このことから着想を得て、運転動作生成領域[30, 34]では、エージェントの軌跡を離散的な動作トークンにトークン化し、自己回帰のためにクロスエントロピー損失に基づくNTP(Next Token Prediction)タスクを採用した研究がある。これらのモデルは、エンコーダ・デコーダのアーキテクチャを利用し続け、連続的なベクトル化された地図と過去の軌跡データをエンコーダでエンコードし、デコーダモジュールのみで離散トークンをデコードする。連続分布回帰法と比較して、NTPの自己回帰パラダイムには以下の利点がある:モデルはステップバイステップの次のトークン予測を採用し、各時間ステップにおけるエージェントの動作間の相互作用をモデル化することを可能にし、モダリティの数は制限されないため、生成タスクの多様性が向上する。

しかし、既存のNTPベースのモーションモデルは、産業用途に重大な影響を与える前述の一般化可能性とスケーラビリティの問題にまだ対処できていない。一般化可能性とは、ゼロショット学習と少数ショット学習により、多様なデータセットで満足のいく結果を得ることを意味し、スケーラビリティとは、[16]で定義されたスケーリング則に従い、データセットサイズやモデルパラメータが増加するにつれて、モデル性能を向上させることを意味する。この不足は主に2つの要因によるものである: 第一に、現在のモデル・アーキテクチャは、限られたデータ規模という制約の下では一般化可能性に欠ける。オープンソースのデータセットは、膨大な運転データを取得するコストが高いため、特定の都市部における運転時間は通常数百時間しかカバーしておらず、知覚や地域差によるドメインギャップが大きい。第二に、単一次元の直列化を伴うタスクとは異なり、モーション生成は、軌跡の時間的次元と、マップとエージェント間の空間的相互作用の両方の直列化を必要とする。これらの課題に取り組むため、本稿ではSMARTモデルを紹介する: 次トークン予測によるスケーラブルなマルチエージェントリアルタイムモーション生成。このモデルは地図データ用のトークナイザーを組み込み、モデルの空間理解度を高めるために、次の道路トークン予測用の自己回帰予測タスクを提案する。その後、GPTスタイルのアプローチを採用し、時系列全体にわたるエージェントの軌跡をトークン化し、デコーダのみの変換モデルを確立する。デコーダのみの変換器により、SMARTは推論中の現在の瞬間に次のフレームの次のトークンを計算することができ、推論ごとに過去のモーショントークンを再エンコードする必要がなくなり、リアルタイム対話型自律走行シミュレーションの推論効率が大幅に向上する。

要約すると、我々のコミュニティへの貢献は以下の通りである: (1)ベクトル化された道路とエージェントの軌跡の両方に対するトークン化スキームを組み込み、次のトークン予測タスクの学習にデコーダのみの変換器を利用する、動き生成のための新しいフレームワークを提案する。このアプローチは、自律走行のための運動生成アルゴリズムの設計に新たな洞察を与える。(2) 運転運動生成の分野では、異なるデータセット間でのモデルのゼロショット汎化性に焦点を当てた先駆的な研究を行っている。注目すべきは、NuPlanデータセットのみで学習したモデルは、これら2つのデータセットの地図領域が重複していないにもかかわらず、WOMDテストデータセットで良好な結果を示したことである。SMARTモデルのスケーラビリティを実証的に検証することで、大規模ファンダメンタルズモデルの魅力的な特性をエミュレートする。(3) SMARTは、生成的Sim Agentsチャレンジのほとんどのメトリクスで最先端の性能を達成し、WOMDリーダーボード[2]で1st位を獲得した。さらに、SMARTのシングルフレーム推論時間は15ms以内であり、自律走行における対話型シミュレーションのリアルタイム要件を満たしている。

2 Related work

2.1 自己回帰型ラージモデルの性質

スケーラビリティとゼロショット汎化 Power-lawスケーリング則[22, 12, 31]は、モデルパラメータ、データセットサイズ、計算資源の成長と機械学習モデルの性能向上の関係を数学的に記述し、いくつかの明確な利点を提供する。

[2] https://waymo.com/open/challenges/2024/sim-agents/

まず、モデルサイズ、データサイズ、計算コストをスケールアップすることで、より大きなモデルの性能を外挿することができる。第二に、スケーリング則は一貫した非飽和的な性能向上を示し、モデル能力を向上させる上での持続的な優位性を裏付けている。ゼロショット生成とは、未知のデータセットから時系列の予測モーションを生成するモデルの能力のことである。ゼロショット生成に関する先行研究[29, 21]では、通常、単一の時系列データセットで学習し、異なるデータセットでテストする。本研究では、SMARTモデルの学習にNuPlanデータセットを、テストにWOMD検証データセットを利用する。自律走行分野における既存の手法[37, 40]は、一般化可能性と解釈可能性を高めるために、意思決定と計画を支援するためにLLMやVLMに依存することが多い。しかし、スケーラビリティとゼロショット汎化性を検証するために、駆動運動場の基礎モデルを直接構築しようとした研究はない。

2.2 連続領域におけるトークナイザー

言語モデル[42, 43]は、テキストのトークン化にバイトペアエンコーディングやWordPieceアルゴリズムに依存している。言語モデルに基づく視覚生成モデル[49, 48]では、2次元画像を1次元トークン列に符号化することも必要である。初期の試み VQVAE [44]は、再構成品質は比較的中程度であったが、画像を離散的なトークンとして表現する能力を実証している。走行運動領域では、MotionLM[34]は、エージェント軌道の連続するウェイポイント間の軸合わせされたデルタの単純な一様量子化を使用した。

2.3 運転モーションの生成

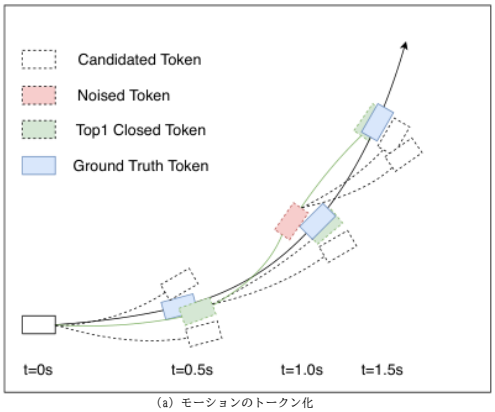
我々の研究は、運転モーション生成における最近の進歩に大きく基づいている。この問題には、連続運動分布回帰[33, 1, 39]、拡散モデル[50, 20]、離散自己回帰モデル[30, 34]など、包括的な生成モデルが適用されている。MotionDiffuser[20]は、複数のエージェントにまたがる将来の軌道の共同分布をモデル化するための拡散ベースの表現手法であり、効率的でトップパフォーマンスのマルチエージェント運動予測のために、単純な予測器設計とPCA圧縮を活用している。これらの拡散に基づくモデルは、個々のエージェントのマルチモーダルな未来軌跡を生成するが、可能性のあるエージェントの動きの限界分布を捉えるだけであり、エージェントの将来の動き間の相互作用をモデル化するものではない。典型的な分布回帰モデルは、将来の運動分布をモデル化するために、ガウス[36]やラプラス[53]のようなパラメトリック連続分布を使用する。これらのモデルの限界は、ガウス混合分布とラプラス混合分布のどちらが将来の状態に対する分布を十分に表現できるかが不確実であることである。さらに、マルチモーダルな未来モーションを生成するために、これらのモデルはしばしば、デコーダモジュールにマルチモーダルなクエリとしてモーションゴール候補[13]や学習可能な潜在埋め込み[45]を組み込む必要があり、その結果、メモリ使用量が多くなり、推論時間が増加する。MotionLM[34]は、自律走行車におけるマルチエージェントの動き予測を言語モデリングタスクとして扱い、複雑な最適化や潜在的なアンカー埋め込みを必要とせず、簡略化された自己回帰プロセスを通じて対話的な軌道を生成する。これに基づき、Trajeglish [30]はマルチエージェントオフライン閉ループシミュレーションを対象としている。

3 Method

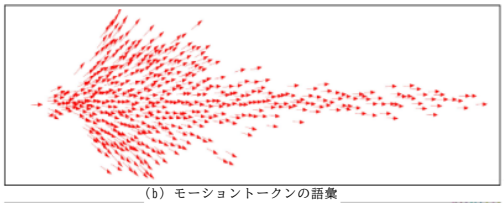
本節では、動的運転シナリオのための自己回帰生成モデルであるSMARTを紹介する。言語運動とエージェント運動はともに逐次的であるが、その表現が異なる。自然言語は有限の語彙からなるのに対し、エージェント運動は連続的な実数値データである。この区別は、語彙の構築やモーションシーケンスのトークン化など、エージェントモーションと道路ベクトルトークナイザーについて、第3.1節で概説したユニークな設計を必要とする。3.2節では、モデルのアーキテクチャを包括的に説明する。第3.3節では、提案モデルが時間シーケンス内のモーショントークンの分布と空間シーケンス内の道路トークンの分布を学習するために設計された学習タスクについて詳しく説明する。

3.1 Tokenization

エージェントモーションのトークン化 VQVAE [44]やVQGAN [10]のような事前に訓練されたトークナイザーを使用して、連続的な特徴を離散的なトークンにエンコードするか、連続的な特徴を正規化する。



(a) モーションのトークン化


(b) モーショントークンの語彙

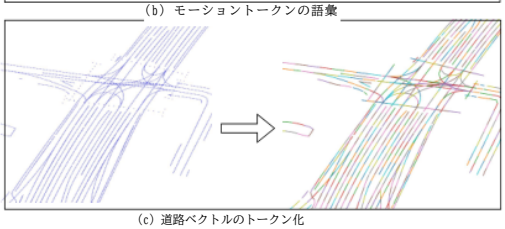

(c) 道路ベクトルのトークン化

図1：(a) 時刻t=0sにおいて、現在の車両状態を参照として、トークン集合内のグランドトゥルースのバウンディングボックスに最も近いトークンを選択する。時刻t=0.5sで、前のステップのマッチしたトークンが次の予測トークンを選択するために使われる。時刻t=1.0sにおいて、ノイズの入ったトークンが、t=1.5sの間、トークンを決定するための参照となる。この反復プロセスは続く。(b) 時間粒度が0.5sに等しいモーショントークン語彙 (c) 元の道路ベクトル特徴は、地図点の連続シーケンスとして表現される。元のマップを複数のセグメントに分割し、それぞれの長さを5m以内に設定し、離散トークンによるマッチングを行う。最終的なマップは、異なる色のセグメントで表現された道路ベクトルトークンで構成される。

の特徴を抽出し、連続値を等間隔で離散スロットに分割する[2, 34]。前者のアプローチでは、潜在的な語彙を確立するために、トークナイザーを訓練するために大量の生データを必要とすることが多い。そうでなければ、トークナイザー自体は訓練前のデータセットに偏ってしまう。我々の研究は、少数のデータサンプルで学習した場合に、モデルが効果的に汎化できるようにすることを目的としているため、SMARTは明示的な軌跡と地図の特徴を離散化することを選択した。具体的には、[30]と同様に、データセット中の全てのエージェントの連続軌跡を、一定の時間間隔t = 0.5sで軌跡集合に分割する。次に、k-disksアルゴリズムを用いて軌跡集合をクラスタリングする。図1(b)に示すように、サンプリングされた軌跡は最終的なエージェントモーショントークン語彙V_aとなる。図1(a)に示すように、青いボックスは、グランドトゥルースの軌跡を離散化した後に得られたトークンを表す。0.5秒間隔で、トークンの語彙の中からトークン候補を探し、そこから適切な(最も近い)トークンを選択して現在の瞬間を表現する。なお、エージェント動作シーケンスのトークン化過程で発生する可能性のあるマッチングエラーを防ぐため、与えられた期間Tにおける連続動作文全体に対して、ローリングマッチングアプローチを実装している。これは、次の時間ステップのトークンが、実際の正しい位置に依存するのではなく、現在マッチしているトークンの位置を参照することによってマッチングされることを意味する。しかし、変換デコーダは段階的に逐次推論を行う必要があるため、このアプローチは必然的に複合エラーによる分布外問題につながる[32]。特に自律走行の分野では、これらの累積誤差が衝突やマップ外事象を引き起こす可能性がある[51]。この問題に対処するため、トークン化処理にノイズを導入し、学習中の分布シフトをモデルがシミュレートできるようにする。具体的には、語彙中のグランドトゥルーストークンに最も近い上位k個のトークンから1個を選択することで、現在マッチしているトークンを摂動する。次に、次の時間ステップで、摂動された車両の状態に基づいてモーショントークンを照合する。このデータ補強法により、分布のずれや累積誤差などの問題を効果的に扱うことができ、生成タスクにおける頑健性が向上する。最後に、エージェントのモーショントークンは$A \in_{R}^{NA \times NT \times FA}$と表され、$N_A$はエージェントの総数を表し、$N_T$は座標、方位、形状を含む特徴サイズ$F_A$の時間ステップ数を表す。

道路ベクトルのトークン化モデルの汎化能力を高めるために、エージェントの動きで行ったのと同様のトークン化処理を道路ベクトルに適用した。各道路ベクトルは、データセットから開始位置と終了位置、長さ、旋回方向、その他のセマンティクスを含む特徴を持つ有向車線セグメントである。道路ネットワークのきめ細かな入力を得るために、すべての道路ベクトルは長さ5m以下のトークンに分割される。

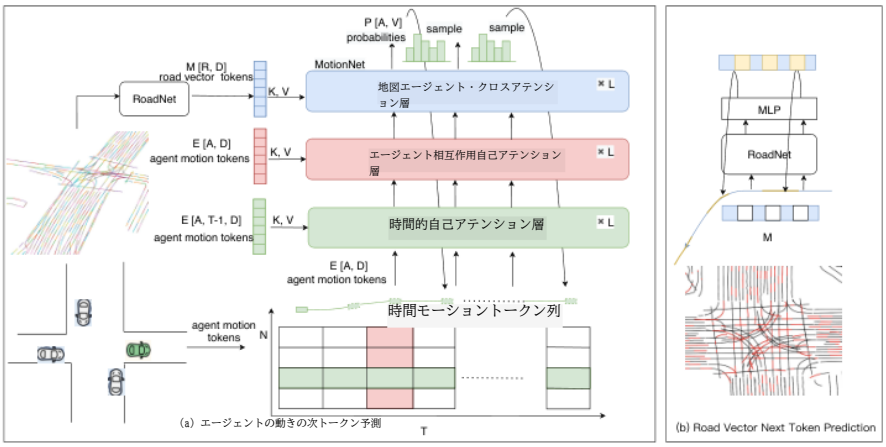


図2:SMARTフレームワークのアーキテクチャ（a）過去のモーショントークン、対話エージェントのモーショントークン、道路トークンの符号化を条件として、マルチエージェントのモーショントークンを予測するデコーダのみの変換器を学習する。モデルは次のモーショントークンを予測するように学習される。(b）提案する道路空間理解トレーニングタスクの説明図。

モーションシーケンスとは異なり、道路文のトークン化処理には時系列依存性がない。図1(c)に示すように、道路文のトークン化は並列に行われ、元の道路ベクトルセグメントをすべて直接トークン化する。道路ベクトルトークンは$R \in R^{N_R \times F_R}$で表され、N_Rは道路ベクトルの総数を表し、F_Aはトークン特徴を表す。

3.2 モデルアーキテクチャ

図2は、SMARTのシンプルで表現力豊かなモデル・アーキテクチャを示している。このモデルは、道路地図符号化のためのエンコーダと、モーショントークン埋め込みに基づいてカテゴリ分布を予測するモーションデコーダから構成される。

RoadNet: 道路トークンエンコーダ道路トークン間の関係をモデル化するために、マルチヘッド自己注意(MHSA)を採用し、その後、更新された道路トークンエンコーディングはモーショントークンデコーディングを支援する。i^{th}番目の道路トークンに対して、その埋め込みr_iからクエリを導出し、隣接トークン$r_j \in R_i$に注目させる:

$$r_{i'} = \text{MHSA}\left(q(r_i),\ k(r_j, \text{RPE}_{ij}), v(r_j, \text{RPE}_{ij})\right), \quad j \in R_i \tag{1}$$

ここで、R_iは道路トークンの近傍集合を表す。地図符号化のための空間認識を取り入れるために、r_jと相対位置埋め込みRPE_{ij} [8]の連結からj^{th}個のキー/値ベクトルを生成する。

MotionNet: 因数分解されたエージェントモーションデコーダ エージェントをエンコードするための一般的な方法は、エージェントの動きの時間的ダイナミクスを優先的にキャプチャし、その後、[35]によって強調されたように、エージェント-マップとエージェント-エージェントの相互作用の統合が続く。因数分解された注意は、時間スケールにまたがる詳細なエージェントとマップの相互作用を効果的に捉える[28]。我々の研究では、時系列に沿った複雑な道路-エージェント関係とエージェント-エージェント関係をデコードするために、マルチヘッドクロスアテンション(MHCA)を持つ因数分解Transformerアーキテクチャを活用する。クエリ中心の方法論[52]と同様に、エージェントのローカル座標フレームを区別するために相対位置埋め込みを利用し、対称的な符号化を可能にする。時間ステップtにおけるi^{th}番目のエージェントを例とする。式2aで示されるように、エージェントの動きトークンの埋め込みe tiから得られるクエリが与えられたとき、時間ステップt − τから時間ステップt − 1までのi^{th}番目のエージェントのトークン埋め込みと、それに対応する相対位置埋め込みとのキーと値に基づいて計算することにより、時間的注意を採用する。

$$e_{i'} = \text{MHSA}\left(q(e_i^t), k(e_i^{t-\tau}, \text{RPE}_i^{t,t-\tau}), v(e_i^{t-\tau}, \text{RPE}_i^{t,t-\tau})\right), \quad 0 < \tau < t \tag{2a}$$

$$e_{i'} = \text{MHCA}\left(q(e_i^t), k(r_j, \text{RPE}_{ij}), v(r_j, \text{RPE}_{ij})\right), \quad j \in N_i \tag{2b}$$

$$e_{i'} = \text{MHSA}\left(q(e_i^t), k(e_j^t, \text{RPE}_{ij}^t), v(e_j^t, \text{RPE}_{ij}^t)\right), \quad j \in N_i \tag{2c}$$

同様に、式2bと式2cにおいて、エージェントマップとエージェントエージェントの注意のキーと値は、それぞれ近傍の道路トークンr_j，$j \in N_i$とエージェントの運動トークン$e\ tj$，$j \in N_i$から導かれ、近傍集合N_iは距離閾値50mで決定される。時間的注意、エージェント-エージェント、エージェント-マップ注意を1つの融合ブロックとして順次積み重ね、そのようなブロックをK回繰り返す。

3.3 時空間ネクストトークン予測

学習段階では、交通シーンの時間的・空間的関係を理解するためにSMARTを学習する。これはRoadNetとMotionNetの2つの次のトークン予測タスクで達成され、2つのタスクの目的の和でモデルが最適化される。

道路ベクトルの次のトークン予測図2(b)に示すように、道路ベクトルNTPタスクは、道路ベクトル入力の空間構造を学習するためにRoadNetを対象とする。エージェントの動きとは異なり、道路ベクトルはシーケンスではなくグラフを形成するため、次のトークン予測タスクを直接適用することは困難である。この問題に対処するため、道路の元のトポロジー情報を抽出し、道路ベクトルトークンを、その前任者と後任者のつながりに基づいて、逐次的な関係でモデル化する。図2(b)に示すように、事前学習NTPタスクでは、道路トポロジーに基づき、先行する道路トークンを用いて後続の道路ベクトルトークンを予測する。このアプローチでは、RoadNetが順序のない道路ベクトル間の接続性と連続性を理解する必要がある。単一のトークン化された道路ベクトル列に対する損失関数は、以下のように定義される。

$$\text{loss}(\gamma) = -\sum_{j=1}^{J}\sum_{i=1}^{V_r}(\mathbf{r}_i^{j+1} == \mathbf{r}_{i^{gt}}^{j+1}) \log(p_\gamma(\mathbf{r}_i^{j+1}|\mathbf{r}_i^{1:j})) \tag{3}$$

ここで、p_γ（r_i^{j+1} $|r_i^{1:j}$）はγでパラメータ化されたRoadNetによって予測されるカテゴリ分布を表し、Jはまだ道路ベクトルトークンに分割されていない完全なポリラインを表し、$r^{1:j}$は前任者の道路トークン埋め込みを表し、r_i^{j+1}は次に予測される道路ベクトルトークンを表す。この損失関数は、RoadNetが直前のトークンから正しい次の道路ベクトルトークンを予測することを学習することを保証し、それによって道路ネットワーク内の空間的連続性と接続性を捕捉する。

モーションネクストトークン予測 Motion NTPタスクは、エージェントのモーションの時間依存性だけでなく、エージェントマップとエージェントエージェント間の空間依存性を理解するために、MotionNetをターゲットとする。SMARTは、グランドトゥルースのトークンラベルの分布と予測分布の間のクロスエントロピーを最小化するように学習される。形式的には、1つのトークン化されたモーション文に対する損失関数は次式で与えられる：

$$loss(\theta) = -\sum_{t=1}^{T}\sum_{i=1}^{V_a}(a_i^{t+1} == a_{i^{gt}}^{t+1}) log(p_\theta(a_i^{t+1}|e_i^{1:t}, r_j)) \tag{4}$$

ここで、p_θ（a_i^{t+1} $|e_i^{1:t}$, m_j）はθでパラメータ化されたモデルによって予測されたカテゴリ分布、$e_{1:t}$ は過去のトークン化されたエージェント運動埋め込み、$a_i^{t+1} \in A$ は次に予測されたエージェント運動トークン、r_j はトークン化された近傍道路ベクトル系列を表す。SMARTは分類[41]によって自己回帰を行うことに注意してください。カテゴリカルな出力分布をオプトすることで、出力分布の構造に制約がなくなり、モデルがマルチモーダルな分布を含む任意の分布を学習できるようになるという重要な利点がある。この柔軟性は、多様なデータセットからのエージェントと道路トークンが異なる出力分布パターンに従う可能性があるため、基本的なモデルにとって特に価値がある。

4 Experiments

SMARTモデルの一般化可能性とスケーラビリティを検証するために、我々は広範な実験を行い、様々なスケールでモデルを訓練した。公式WOMD Sim Agents Challenge (WOSAC)では、WOMDデータセットのみで学習させたSMART 700万パラメータ(7M)モデルを採用した。同時に、SMART 7Mモデルは汎化実験やアブレーション研究にも利用された。

表1:WOMD 2023 Sim Agentsベンチマークにおける最先端モデルとの比較

Method	リアリズム・メタ指標	Kinematic metrics↑	Interactive metrics↑	地図ベースの測定基準	minADE↓
SMART 7M	**0.6587**	0.4190	**0.8014**	**0.8523**	1.7453
Trajeglish[30]	0.6451	0.4166	0.7845	0.8216	1.5712
MVTE[46]	0.6448	0.4202	0.7666	0.8387	1.6770
VPD-PRIOR	0.6315	**0.4261**	0.7233	0.8330	1.3400
QCNeXt[53]	0.4538	0.3109	0.5654	0.5051	**1.0830**
MultiPath[45]	0.4766	0.1792	0.6380	0.6866	2.0517

スケールロー実験では、追加のデータセットを統合し、複数のスケールのモデルで学習させた。すべての実験において、テストデータセットはWOMDの分割検証データセットを採用した。SMARTアーキテクチャの詳細なハイパーパラメータはセクションA.1にある。以下のセクションでは、セクション4.1でWOSACベンチマーク[26]でSMARTが生成したロールアウトの結果を示す。SMARTの一般化可能性とスケーラビリティの評価は、それぞれ4.2節と4.3節で詳述する。最後に、4.4節で我々の設計手法のアブレーション分析を行う。

4.1 モーション生成タスクの比較

性能比較提案するSMARTを、拡散モデル[15]、連続分布回帰モデル[46, 36]、次トークン自己回帰モデル[30]などの既存の運動生成アプローチと比較する。Sim Agentsのチャレンジメトリクスは2回変更されたため、より広範に以前の方法論と比較するために、WOMD Sim Agents 2023と2024 Benchmark[26]の両方を使用して、我々のモデルの性能をテストする。表1、表2に示すように、SMARTは最良のRealism Meta指標だけでなく、高い予測精度を達成している。SMARTの地図と動きに対するモデリングアプローチにより、先行研究よりも効果的にデータ内の行動分布を学習することができる。注目すべきは、SMART-zeroshotはNuPlanデータセットのみで学習され、Waymoテストセットで直接推論されたモデルである。表2に示すように、MVTEに近い性能を達成している。詳細な比較については、A.2 を参照されたい。

表2:WOMD 2024 Sim Agentsベンチマークにおける最先端モデルとの比較

Method	Realism Meta metric↑	Kinematic metrics↑	Interactive metrics↑	Map-based metrics↑	minADE↓
SMART 101M	**0.7614**	**0.4786**	**0.8066**	**0.8648**	**1.3728**
SMART 7M	0.7591	0.4759	0.8039	0.8632	1.4062
BehaviorGPT	0.7473	0.4333	0.7997	0.8593	1.4147
GUMP	0.7431	0.4780	0.7887	0.8359	1.6041
MVTE	0.7302	0.4503	0.7706	0.8381	1.6770
SMART-zeroshot	0.7210	0.4311	0.7806	0.8099	2.5703
VBD	0.7200	0.4169	0.7819	0.8137	1.4743
TrafficBOTv1.5	0.6988	0.4304	0.7114	0.8360	1.8825
congniBOTv1.5	0.6288	0.3293	0.7129	0.6918	-

効率比較SMARTは、マルチエージェントモーション生成においても顕著な速度を示している。これまでのエンコーダ・デコーダモデル[34, 36]は、マルチモーダルなモーションを生成するために、デコーダモジュールに複数のクエリ埋め込みを必要とするため、高い計算コストに悩まされている。デコーダのみの変換器アーキテクチャの利点を生かし、SMARTは推論中の現在の瞬間に次のフレームの次のトークンを計算するだけでよく、過去のモーショントークンを再エンコードする必要はない。過去の観測時間地平で計算されたトークン埋め込みを再利用することで、エージェントモーションデコーダの複雑さは$O(N_A\ N_T) + O(N_A\ N_R) + O(N_A^2)$に低減される。一方、[24]のようなエンコーダ・デコーダモデルでは、エンコーダモジュールの計算負荷の他に、軌跡のマルチモダリティを生成するために$O(N_A^2\ N_M) + O(N_A\ N_M\ N_R)$の追加計算が必要であり、ここで$N_M$はモダリティの数を表す。

SMARTの平均シングルステップ推論時間は、マップトークンとエージェントモーショントークンの数に影響され、5〜20msの間で変動し、10ms以下で平均化される。このように、自律走行におけるインタラクティブなリアルタイムオンラインシミュレーションの現在のニーズを大幅に満たしている。

4.2 一般化

異なるデータセットでのゼロショット汎化ゼロショット生成とは、異なるデータセットからの時系列に対してモーションを生成するモデルの能力のことである。本研究では、NuPlanデータセットの学習データを用いてSMARTモデルを学習し、WOMD検証データセットのテストデータを用いている。表3に示すように、SMART*は依然として全体的な指標において良好な性能を達成している。異なるデータセット間で、エージェントの位置と方位に関する較正されたグランドトゥルース値の精度に大きな違いがあるため、エージェントの運動学的メトリクスに大きなギャップがあり、その結果スコアが低くなる可能性がある。しかし、SMART*はエージェントの相互作用とドライバブルマップの指標において優れた汎化性を示した。2つのデータセットのサイズに大きな違いがないので、SMARTモデルは少数のデータ学習に基づいて良好な汎化能力を持つことができることは特筆に値する。

表3:異なるデータセットでのゼロショット汎化。SMARTはWOMDのみで学習したモデルを示す。SMART*はNuPlanデータセットのみで学習したモデルを示す。SMART**は、SMART*モデルに基づくWOMDの初期学習率0.0001で、1エポックの微調整を行った後のモデルを示す。

Method	運動学的指標	Interactive metrics↑	Map-based metrics↑	minADE ↓
SMART	0.4537	0.8034	0.8514	1.5127
SMART*	0.4161	0.7853	0.7970	2.3041
SMART**	0.4310	0.8087	0.8559	1.5671

図3に示すような複数のマップシナリオは、WOMDデータセットにのみ存在し、NuPlanデータセットには存在しない。ネットワークアーキテクチャやチューニングパラメータを変更することなく、NuPlanのみで学習したSMARTは、これらのシナリオで適切な結果を達成し、SMARTの汎化能力を立証している。

4.3 Scalability

先行研究[22, 42]では、大規模言語モデル(LLM)をスケールアップすると、テスト損失Lが予測通りに減少することが立証されている。この傾向は、パラメータ数N、学習トークンTとべき乗則に従って相関する:

$$log(L) = \beta log(X) + \alpha \tag{5}$$

ここで、X は N, T のどれでもよい。指数αはべき乗則の滑らかさを反映し、Lは既約損失で正規化した還元可能損失を表す。スケーリング則を検証するためのデータソースはA.3に詳述されている。全体として、2.2Mのシナリオ(または0.5sエージェントモーショントークン化の下での1Bモーショントークン)を含むトレーニングセットで、1Mから100Mのパラメータまでの4つのサイズにわたってモデルをトレーニングした。

モデルパラメータを用いたスケーリング則モデルサイズの増加に伴うテスト損失の傾向を調査する。最終的なテストクロスエントロピー損失Lを10万件のトラフィックシナリオの検証セットで評価した。結果は図4にプロットされており、モデルサイズNの関数として、損失Lの明確なべき乗スケーリング傾向が観察された。べき乗スケーリング則は次のように表すことができる:

$$log(L) = -0.157log(X) + 1.52 \tag{6}$$

これらの結果は、SMARTの強力なスケーラビリティを検証し、モデル性能がデータセットサイズによってどのようにスケールするかについての貴重な洞察を提供する。

4.4 Ablation

本研究では、SMARTの各コンポーネントの有効性を検証することを目的とする。結果を表4に示す。初期モデル(M 1と表記)は、Sec.3.2で描かれたアーキテクチャを用い、エージェントのトークン化のみを用いて構築されている。道路ベクトルの状態を離散的なトークンにトークン化するM 2に道路ベクトルトークン化を導入することで、汎化能力がM 1より著しく向上する。

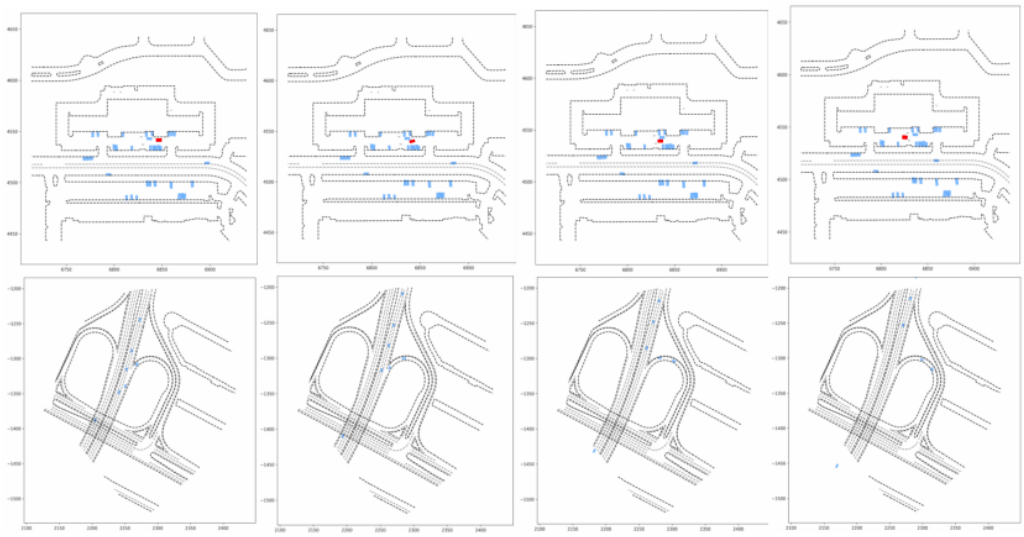
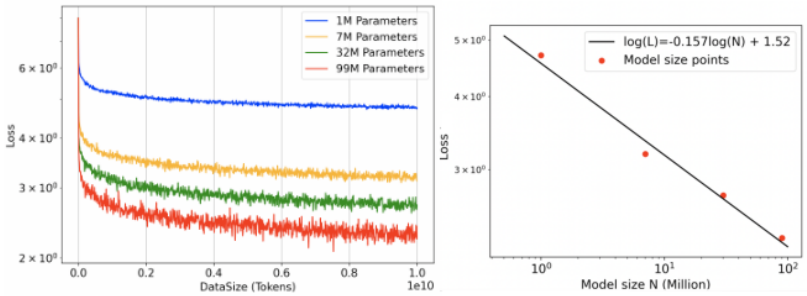


図3: テストセットから2つの代表的なシナリオに対する閉ループ計画の定性的結果。各シナリオ(各行)は8秒間で、2秒間隔で4つのスナップショットを取る。SMARTはシナリオ内のすべてのエージェントを制御する。1行目は駐車場エリアを表している。写真中の赤い車は、駐車場の前方にある静止した車を迂回することを効果的に完了した。2行目は、ランプゾーン内の曲率Uターンが大きく、ランプの右車線の交通流がSMARTの制御下でランプ出口の挙動を完了したシーンである。その他の動画については補足資料を参照されたい。



(a)10億トークンで学習したスケールの異なるモデルの損失削減曲線

(b) SMART変換モデルサイズによるスケーリング則、べき乗則によるフィット(実線)と方程式(凡例)。

図4:データセットサイズの制限のため、合計10億トークンに対して1Mから101Mまでの複数のスケールでモデルを学習させた。(a) 異なるモデルの学習損失 (b) 軸はすべて対数スケール。べき乗スケーリング則は実線で表すことができる。指数β=-0.157は、SMARTモデルをスケールアップしたときのテスト損失Lが滑らかに減少することを示唆している。

モデルM 1とM 2を比較すると、WOMDデータセットのみで学習した場合、道路ベクトルのトークン化により、全体的なメトリクスがある程度減少することがわかる。我々は、離散化された地図トークンは、道路に関するいくつかの細かい幾何学的情報を失う可能性があると推測している。M 4はノイズの多いエージェントの動きのトークン化を組み込んでおり、推論中の累積誤差や分布のずれに対処するように設計されている。この修正により、相互作用メトリックとマップベースメトリックの両方が強化される。

5 Conclusions

本論文では、GPTスタイルのフレームワークでデコーダのみの変換器アーキテクチャで処理された、ベクトル化された地図とエージェントの軌跡データを活用する自律走行運動生成のための新しいパラダイムであるSMARTを紹介した。

表4:SMARTの各コンポーネントに関するアブレーション研究。実験結果はWOMD検証セットに基づく。”RVT ”は道路ベクトルのトークン化、”RVNTP ”は道路ベクトルの次のトークン予測、”NAT ”はノイズ付きエージェントのトークン化、”NRVT ”はノイズ付き道路ベクトルのトークン化を示す。

SMART					Train on WOMD			Train on NuPlan		
Model Number	RVT	NAT	NRVT	RVNTP	kinematics	interactive	map	kinematics	interactive	map
M_1					**0.459**	**0.827**	**0.857**	0.376	0.593	0.603
M_2	√				0.434	0.807	0.840	0.389	0.696	0.724
M_3	√	√			0.448	0.809	0.848	0.413	0.750	0.743
M_4	√	√	√		0.437	0.801	0.837	0.411	0.747	0.741
M_5	√	√		√	0.453	0.813	0.853	0.413	0.780	0.785
M_6	√	√	√	√	0.453	0.803	0.851	**0.416**	**0.785**	**0.797**

SMARTは、スケーラビリティとゼロショット汎化という2つの重要な特性をエミュレートしており、これらは大規模モデルの進化に不可欠である。我々は、我々の発見と全てのコードの公開が、自律走行分野における運動生成のためのモデルのさらなる探求と開発を促し、最終的にはより信頼性の高い自律走行システムに貢献すると考えている。

限界点本研究では、主に学習パラダイムの設計に焦点を当て、離散トークン語彙の比較的単純な設計を維持する。我々は、高度なトークナイザー[25]やサンプリング技術でSMARTを反復することで、さらに性能を向上させることができると考えている。複数のデータセットから学習データを収集したが、モデルのスケーラビリティを検証する際、データセットサイズに制限があり、最大スケールが1億パラメータのモデルに限定している。この研究の汎化とスケーリング則に焦点を当てると、エージェントの動きトークンの時間粒度やトークン語彙のサイズなど、多くのハイパーパラメータアブレーション実験が検証される必要がある。運動生成モデルとして、SMARTが計画や予測タスクに移行する能力はまだ検証される必要があり、これが今後の研究の最優先課題である。

謝辞:匿名査読者、エリアチェア、プログラム委員の貴重な示唆に感謝する。また、本研究の質を大幅に向上させることができた。また、Yue Gong氏、Shuxiang Lu氏との思慮深い議論に感謝する。著者 Wei Wu、Xiaoxin、Ziyan はこの研究に等しく貢献した。Wei Wuはプロジェクトを主導し、資金援助を行い、XiaoxinとZiyanはアルゴリズム設計、実装、モデルトレーニング、原稿執筆に重点を置いた。

References

[1] Elmira Amirloo, Amir Rasouli, Peter Lakner, Mohsen Rohani, and Jun Luo. Latentformer: Multi-agent transformer-based interaction modeling and trajectory prediction. *arXiv preprint arXiv:2203.01880*, 2022.

[2] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

[3] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

[4] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv preprint arXiv:1910.05449*, 2019.

[5] Yongli Chen, Shen Li, Xiaolin Tang, Kai Yang, Dongpu Cao, and Xianke Lin. Interaction-aware decision making for autonomous vehicles. *IEEE Transactions on Transportation Electrification*, 2023.

[6] Jie Cheng, Yingbing Chen, Xiaodong Mei, Bowen Yang, Bo Li, and Ming Liu. Rethinking imitation-based planner for autonomous driving. *arXiv preprint arXiv:2309.10443*, 2023.

[7] Alexander Cui, Sergio Casas, Abbas Sadat, Renjie Liao, and Raquel Urtasun. Lookout: Diverse multi-future prediction and planning for self-driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16107–16116, 2021.

[8] Alexander Cui, Sergio Casas, Kelvin Wong, Simon Suo, and Raquel Urtasun. Gorela: Go relative for viewpoint-invariant motion forecasting. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7801–7807. IEEE, 2023.

[9] Henggang Cui, Vladan Radosavljevic, Fang-Chieh Chou, Tsung-Han Lin, Thi Nguyen, Tzu-Kuo Huang, Jeff Schneider, and Nemanja Djuric. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2090–2096. IEEE, 2019.

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[11] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021.

[12] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

[13] Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15303–15312, 2021.

[14] Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, et al. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *Advances in Neural Information Processing Systems*, 36, 2024.

[15] Zhiming Guo, Xing Gao, Jianlan Zhou, Xinyu Cai, and Botian Shi. Scenedm: Scene-level multi-agent trajectory generation with consistent diffusion models. *arXiv preprint arXiv:2311.15736*, 2023.

[16] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[17] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023.

[18] 黄志宇、ピーター・カルクス、ボリス・イワノビッチ、陳玉暁、マルコ・パヴォネ、チェン・ルヴ。Dtpp: 自律走行における樹木政策計画のための微分可能な共同条件付き予測とコスト評価。arXivプレプリント arXiv:2310.05885, 2023.

[19] 黄志宇、劉浩晨、陳呂。ゲームフォーマー: ゲーム理論に基づく自律走行のための変換器ベースの対話的予測・計画のモデリングと学習。arXivプレプリントarXiv:2303.05760, 2023.

[20] Chiyu Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, Dragomir Anguelov, et al: 拡散を用いた制御可能なマルチエージェント動作予測。コンピュータビジョンとパターン認識に関するIEEE/CVF会議予稿集, ページ9644-9653, 2023.

[21] 金暁永、パク・ヨンスク、ダニエル・マディックス、王浩、王玉陽。アテンションシェアリングによる時系列予測のためのドメイン適応。機械学習国際会議、10280-10297ページ。PMLR, 2022.

[22]ジャレッド・カプラン、サム・マッカンドリッシュ、トム・ヘニガン、トム・B・ブラウン、ベンジャミン・チェス、リウォン・チャイルド、スコット・グレイ、アレック・ラドフォード、ジェフリー・ウー、ダリオ・アモデイ。ニューラル言語モデルのスケーリング則arXivプレプリント arXiv:2001.08361, 2020.

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[24] Osama Makansi, Eddy Ilg, Ozgun Cicek, and Thomas Brox. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7144–7153, 2019.

[25] ファビアン・メンツァー、デビッド・ミンネン、エイリクル・アグストソン、マイケル・ツチャネン。有限スカラー量子化: Vq-vaeをシンプルにする。arXivプレプリント arXiv:2309.15505, 2023.

[26]ニコ・モンタリ、ジョン・ランバート、ポール・ムーギン、アレックス・クフラー、ニコラス・ラインハート、ミシェル・リー、コール・グリーノ、トリスタン・エムリヒ、エディ・ヤン、シモン・ホワイトソン、他、ウェイモ・オープン・シム・エージェント・チャレンジ。神経情報処理システムの進歩, 36, 2024.

[27] Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S Refaat, and Benjamin Sapp. Wayformer: シンプルで効率的な注意ネットワークによる運動予測。2023年IEEEロボティクスとオートメーション国際会議(ICRA)、2980-2987ページ。IEEE, 2023.

[28] Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, et al: 複数のエージェントの軌跡を予測するための統一的なアーキテクチャ。arXivプレプリント arXiv:2106.08417, 2021.

[29] ベルナルド・ペレス・オロスコ、スティーブン・J・ロバーツ。順序回帰リカレントニューラルネットワークによるゼロショットおよび少数ショットの時系列予測。arXivプレプリントarXiv:2003.12162, 2020.

[30] Jonah Philion, Xue Bin Peng, and Sanja Fidler. Trajeglish: Learning the language of driving scenarios. *arXiv preprint arXiv:2312.04535*, 2023.

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[32] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models, 2023.

[33] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 683–700. Springer, 2020.

[34] Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8579–8590, 2023.

[35] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543, 2022.

[36] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying. *arXiv preprint arXiv:2306.17770*, 2023.

[37] 志茂浩、Katrin Renz、Kashyap Chitta、Li Chen、Hanxue Zhang、Chengen Xie、Ping Luo、Andreas Geiger、Hongyang Li.Drivelm: グラフ視覚質問応答による運転。arXivプレプリント arXiv:2312.14150, 2023.

[38] 宋浩蘭、丁文超、陳玉宣、沈少傑、王マイケル・ユー、陳奇峰。Pip: 自律走行のための計画情報に基づく軌道予測。コンピュータビジョン-ECCV 2020: 第16回ヨーロッパ会議、グラスゴー、英国、2020年8月23日～28日、会議録、パートXXI 16、ページ598～614。Springer, 2020.

[39] サイモン・スオ、セバスチャン・レガラード、セルジオ・カサス、ラクエル・ウルタスン。Trafficsim: 現実的なマルチエージェントの行動をシミュレートするための学習。コンピュータビジョンとパターン認識に関するIEEE/CVF会議予稿集, ページ10400-10409, 2021.

[40] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

[41] Luís Torgo and João Gama. Regression by classification. In *Advances in Artificial Intelligence: 13th Brazilian Symposium on Artificial Intelligence, SBIA'96 Curitiba, Brazil, October 23–25, 1996 Proceedings 13*, pages 51–60. Springer, 1996.

[42] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al: オープンで効率的な基礎言語モデル。arXiv プレプリント arXiv:2302.13971, 2023.

[43] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[44] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[45] Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7814–7821. IEEE, 2022.

[46] Yu Wang, Tiebiao Zhao, and Fan Yi. Multiverse transformer: 1st place solution for waymo open sim agents challenge 2023. *arXiv preprint arXiv:2306.11868*, 2023.

[47] ベンジャミン・ウィルソン、ウィリアム・チー、タンメイ・アガルワル、ジョン・ランバート、ジャギート・シン、シッデシュ・カンデルワル、ボーエン・パン、ラトネシュ・クマール、アンドリュー・ハートネット、ジョニー・ケースモデル・ポンテス、他 Argoverse 2: 自動運転の知覚と予測のための次世代データセット。arXivプレプリント arXiv:2301.00493, 2023.

[48] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[49] リュ・リジュン、ホセ・レズマ、ニテシュ・B・グンダヴァラプ、ルカ・ヴェルサリ、ゾーン・キヒョク、デヴィッド・ミネン、チョン・ヨン、ヴィグネシュ・ビロドカー、アグリム・グプタ、グ・シュイエ、アレクサンダー・G・ハウプトマン、ゴン・ボチン、ヤン・ミンヒョン、イルファン・エッサ、デヴィッド・A・ロス、ル・ジャン。言語モデルは拡散に勝る - トークナイザーは視覚生成の鍵である, 2024.

[50] 周紫源、Davis Rempe、徐丹飛、陳玉暁、Sushant Veer、Tong Che、Baishakhi Ray、Marco Pavone。制御可能な交通シミュレーションのためのガイド付き条件付き拡散。2023年IEEEロボティクスとオートメーション国際会議(ICRA)、3560-3566ページ。IEEE, 2023.

[51] 周金雲、王瑞、劉旭、姜義飛、江秀、田嘉明、苗景浩、宋世宇。フィードバック合成器と微分可能ラスタライズを用いた自律走行のための模倣学習の探求。2021年IEEE/RSJ知能ロボットとシステム国際会議(IROS)、1450-1457ページ、2021年。

[52] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17863–17873, 2023.

[53] Zikang Zhou, Zihao Wen, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Qcnext: A next-generation framework for joint multi-agent trajectory prediction. *arXiv preprint arXiv:2306.10508*, 2023.

A Appendix

A.1 実装とシミュレーション推論

アーキテクチャの詳細表5は、我々の実装で使用された様々なモデルのハイパーパラメータをまとめたものである。3つのカテゴリ(車両、歩行者、自転車)の将来の動きを生成するために、各カテゴリが独自のモーショントークン語彙を持つ単一のモデルを訓練する。入力道路トークン特徴量には、各道路トークン点の位置、各点における道路トークンの方向、各道路トークンのタイプの3種類の情報が含まれている。各デコーダ層の予測ヘッドには3層のMLPを使用し、モデルの重みは異なるデコーダ層で共有されない。

表5:異なるSMARTモデルのハイパーパラメータ

Modules	Hyperparameters	Values "スマート1M	スマート7M	スマート26M	スマート101M"
RoadNet	自己注意層の数 Road token embeddings Road token vocabulary Road token attention radius	1 64 1024 10	3 128 1024 10	1 256 1024 10	3 512 1024 10
MotionNet	時間的注意層の数 エージェント-エージェント注意層の数 地図-エージェント注意層の数 注意ヘッドの数 注意ヘッドの次元	1 2 2 8 8	6 6 6 8 16	6 6 6 8 32	6 6 6 8 64
	Feature dimension of Agent token embeddings	64	128	256	512
	Size of motion token vocabulary	512	1024	1024	2048
SMART	Total parameters	1.0M	7.2M	26.9M	101.0M

トレーニングの詳細 AdamWオプティマイザ[23]を用いて、3つのエージェントタイプすべてについてシミュレーションモデルをエンドツーエンドでトレーニングする。ドロップアウト率、ウェイト減衰率ともに0.1に設定。学習率はコサインアニーリングスケジューラを用いて0.0002から0まで減衰させる。トレーニングには、シーン内のすべての車両が含まれる。バッチサイズは 4 に設定され、最大 GPU メモリ使用量は 30GB である。

WOSACの推論テストセットは44,920シーンからなり、各シーンはエージェントグループの32のシミュレーションを生成するために、モデル推論を32×T回実行する必要がある。モデル推論中、各シミュレーションステップは次のトークンの分類された分布を生成する。次のトークンのサンプリングには2つの選択肢がある:最尤トークンを選択するか、再分配された確率で上位k個のモーショントークンの中からサンプリングする。最初のアプローチは、正確ではあるが、より多様な世代をもたらす傾向がある。逆に、top-kモーショントークンを選択すると、多様性は促進されるが、エラーが複合化し、非現実的な運動学的モーションやドリフトを伴う軌道が生成される可能性がある。現実性と多様性のバランスをとるため、シミュレーション中の各ステップでトップ5のサンプリングを使用する。ロールアウトの動画は、プロジェクトページや補足資料で見ることができる。各シナリオについて、SMARTモデルはシーン内のすべてのエージェントを直接制御する。本稿では、モデルの汎化性とスケーラビリティに焦点を当てるため、詳細なトリックを広範に検討することなく、特定のシーン生成において良好な結果を得ることができた。

A.2 WOSACリーダーボードにおける詳細な比較

表6:尤度を表すWOMDのテスト分割におけるコンポーネントごとのメトリック結果。WOSC 評価指標の計算の更新により、より広範な比較のために、2023 年リーダーボードの複合指標で手法をランク付けしている。最新のWOSCについては、2024年リーダーボードの更新を直接参照し、詳細な比較を行ってください。

Method	KINEMATIC				INTERACTIVE			MAP		minADE↓
	LINEAL SPEED↑	LINEAR ACCEL↑	ANG. SPEED↑	ANG. ACCEL↑	DIST TO OBJ.↑	COLLISION ↑	TTC ↑	DIST TO ROAD↑	OFF ROAD↑	
WAYFORMER	0.202	0.144	0.248	0.312	0.192	0.449	0.766	0.379	0.305	6.823
SBTA-ADIA	0.317	0.174	0.478	0.463	0.265	0.337	0.770	0.557	0.483	3.611
CAD	0.346	0.252	0.432	0.311	0.33	0.311	0.789	0.637	0.539	2.314
JOINT-MULTIPATH++	0.431	0.230	0.019	0.035	0.349	0.485	0.811	0.637	0.613	2.051
MTR+++	0.411	0.106	0.483	0.436	0.345	0.414	0.796	0.654	0.577	1.681
QCNeXt	**0.477**	0.242	0.325	0.198	0.375	0.324	0.756	0.609	0.360	**1.083**
MVTE	0.442	0.221	0.535	0.481	0.382	0.450	0.832	0.664	0.640	1.677
Trajeglish	0.450	0.192	**0.538**	0.485	**0.387**	0.922	**0.836**	**0.659**	0.886	1.571
SMART 7M	0.363	**0.296**	0.423	**0.564**	0.376	**0.963**	0.832	**0.659**	**0.936**	1.749

Waymo Open Sim Agents Challenge (WOSAC)は、自律走行車のためのシミュレーションエージェントの開発と評価を進めることを目的とした重要なイニシアチブである。この課題は、Waymo Open Motion Dataset (WOMD)を活用し、最先端のオフボード知覚システムによって生成された忠実度の高い物体行動と形状を提供する。参加者は、エージェントの行動と相互作用のリアリズムを確保するために、クローズドループ評価に焦点を当て、最大128のエージェントを含むシナリオをシミュレートすることが要求される。評価フレームワークは、実世界の運転データにマッチする現実的で多様な行動を生成するシミュレーションエージェントの性能を評価するために、運動学的特徴、相互作用に基づく特徴、マップに基づく特徴を含む様々なメトリクスを採用している。WOSACは9つの測定値に対して3つのメトリクスを計算する:運動学的メトリクス(線速度、直線加速度、角速度、角加速度の大きさ)、物体相互作用メトリクス(最も近い物体までの距離、衝突、衝突までの時間)、地図ベースのメトリクス(道路端までの距離、道路逸脱)。

表6に示すベンチマーク比較では、我々のチームが開発したSMART 7M法は、複数の指標において優れた性能を示し、特に対話型指標や安全関連指標において優れていることがわかる。注目すべきは、SMART 7Mが角加速度、最近接物体までの距離、衝突回避、オフロードメトリクスで最高のスコアを達成したことで、複雑な運転シナリオにおける有効性が強調された。これらの結果は、安全性と信頼性の確保におけるSMART 7Mの頑健性を浮き彫りにし、動的で潜在的に危険な交通状況を管理する高度な能力を、他の評価手法よりも効果的に示している。この性能は、SMARTモデルが計画タスクに適用できる可能性も示唆している。

A.3 追加アブレーション研究

異なるアーキテクチャのスケーラビリティと汎化性の比較本節では、本論文で提案するアーキテクチャとMVTEモデルを比較する実験を行う。MTRから導かれるMVTEモデル[3]は、連続分布回帰モデルを表す。実験結果

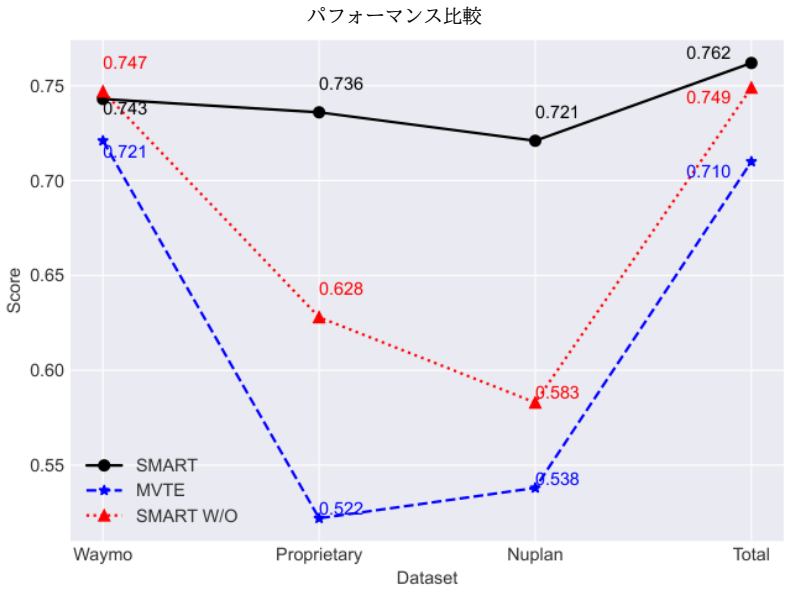


図5:SMART w/oは、本稿で提案する道路ベクトルのトークン化とノイズのトリックを用いないSMARTモデルを指す。実験の公平性を確保するため、すべてのモデルパラメータは90-100Mに調整した。モデルは様々なデータセットで学習され、WOMD検証データセットでのみ検証された。

図5に示す結果は、分布回帰に基づくモデルは、異なるデータセット間で汎化能力が低いことを示している。他のデータセットからのインクリメンタルデータで学習したモデルは、WOMDのみで学習したモデルよりも全体的に性能が悪かった。

[3]Since MVTE does not have open-source code, we reproduced the results by relying on the MTR model

興味深い現象は、我々のプライベートデータセットはNuPlanデータセットよりも多くのデータを含んでいるが、それに対して学習したMVTEの性能はNuPlanデータセットよりも劣っていることである。これは、分布回帰に基づくパラダイムが、モデルをデータセットに過剰適合させる可能性が高いことを示唆している。SMART w/oの結果から、モデルの汎化性能は低いが、インクリメンタルデータの効果により、単一の学習データセットと比較して性能が向上することがわかる。以上の実験から、離散トークン化はデータセットギャップをなくすのに非常に有効な方法であると推察される。さらに、クロスエントロピー分類損失に基づく自己回帰モデルは、軌跡生成モデルのスケーラビリティの鍵であり、大規模言語モデル(LLM)が有意なスケーリング能力を持つ理由と一致する。

異なるトークナイザーの比較 Trajeglish[30]では、様々な離散化トークナイザーの詳細な比較が行われている。本論文の3.1節で紹介したように、最終的にトークン語彙の構築にはk-disksアプローチを採用した。我々の研究以前には、潜在トークン化手法[44]を用いてエージェントと道路の動きトークンの語彙を構築しようとした研究はなかった。そこで、視覚領域のVQ-VAEアプローチを利用して、モーショントークンの潜在的な自動符号化を行い、このトークナイザーと本稿で選択した手法との比較を行った。

表7:異なるトークナイザーの比較。実験結果はSMART 7Mに基づく。

	Train on WOMD			Train on NuPlan		
Tokenizer	Kinematics	Interactive	Map	Kinematics	Interactive	Map
VQ-VAE	0.461	0.810	0.853	0.376	0.687	0.703
K-disks	0.453	0.803	0.851	0.416	0.785	0.797

表7の結果から、VQ-VAEはk-disksと比較して、単一のデータセットでより良い性能を発揮することがわかる。具体的には、両手法とも対話型メトリクスとマップベースメトリクスでは同様の結果を得るが、VQ-VAEは運動学的メトリクスではk-diskを上回る。k-disksアプローチは離散化の際にきめ細かい軌跡情報を失うが、VQ-VAEは軌跡を再構成する際にデータセットの真の分布によく適合する。しかし、ゼロショット汎化における2つの手法の性能を比較すると、k-disksはVQ-VAEを大幅に上回る。我々は、モーションとロードトークンの語彙を構築するVQ-VAEトークナイザーのトレーニング中に、トークナイザーがトレーニングデータセットをすでに記憶していたか、オーバーフィットしていた可能性があると推測している。したがって、VQ-VAEアプローチを用いてより良い汎化性能を達成するためには、大規模なデータセットでVQ-VAEトークナイザーを事前学習させることが不可欠である。

言語モデルの場合、大規模で多様なデータセットを得ることは比較的容易である。対照的に、自律走行モーション領域は、同等のサイズと多様性のデータソースを欠いている。より大きなデータセットでスケーリング則を検証するために、Waymo、Nuplan、そして我々の独自のデータセットからのデータを統合した。スケーリング則を検証するためだけに、独自のデータセットを導入した。WOSCリーダーボードの評価には、Waymoデータセットのみを使用した。汎化実験やその他のアブレーション実験では、NuplanとWaymoのオープンソースデータセットの両方を利用し、広く利用可能なデータセットへのアクセスを提供することで、実験の再現性を容易にした。以下の表8は、各データセットのシナリオ数、継続時間、総モーショントークン数をまとめたものである。

表9の結果は、様々なメトリクスにおいて、異なるパラメータスケールを持つSMARTモデルの性能を強調している。モデル規模がSMART 1MからSMART 101Mに増加するにつれて、対話型メトリクスとマップベースメトリクスの両方が大幅に改善される。これは、より大きなモデルが相互作用を捉え、マップベースのコンテキストを理解することに優れており、これらの分野での性能向上につながることを示している。しかし、運動学的指標は最小限のばらつきしか示さない。

Table 8: Data sources

Dataset	Scene Count	Single Scenario Duration	総モーショントークン数
Nuplan	30w	10s	0.13B
Waymo	48w	9s	0.18B
Proprietary	150w	11s	0.68B
Total	228w	-	1B

表9:スケールを変えたSMARTモデルの比較。学習時間とは、データセット全体を用いてモデルが収束するまでに必要な時間を指す。推論時間とは、1つのフレームに対して次のトークンを予測するのにかかる時間を指す。

Method	Kinematic metrics↑	Interactive metrics↑	Map-based metrics↑	Training time	Average inference time
SMART 1M	0.423	0.782	0.835	8hours	10.30ms
SMART 7M	0.436	0.809	0.852	23hours	17.21ms
SMART 26M	0.442	0.817	0.864	3days	25.94ms
SMART 101M	**0.457**	**0.819**	**0.872**	1week	46.58ms

さらに、学習時間と平均推論時間は、モデルの性能と計算コストのトレードオフを反映して、モデルが大きくなるにつれて大幅に増加する。50,000ステップごとに検証を行う。5回の連続検証の後、有意な損失削減やメトリックの改善が見られない場合、モデルは収束したとみなされる。学習と推論の時間は、32個のNVIDIA TESLA V100 GPUで測定した。