VADv2: 確率的プランニングによるエンドツーエンドのベクトル化自律走行

Shaoyu Chen[1⋆], Bo Jiang[1⋆], Hao Gao[1], Bencheng Liao[1], Qing Xu[2],
Qian Zhang[2], Chang Huang[2], Wenyu Liu[1], Xinggang Wang[1,✉]
[1] Huazhong University of Science & Technology [2] Horizon Robotics
{shaoyuchen, bjiang, g_hao, bcliao, liuwy, xgwang}@hust.edu.cn
{qing.xu, qian01.zhang, chang.huang}@horizon.cc
https://github.com/hustvl/VAD
https://hgao-cv.github.io/VADv2

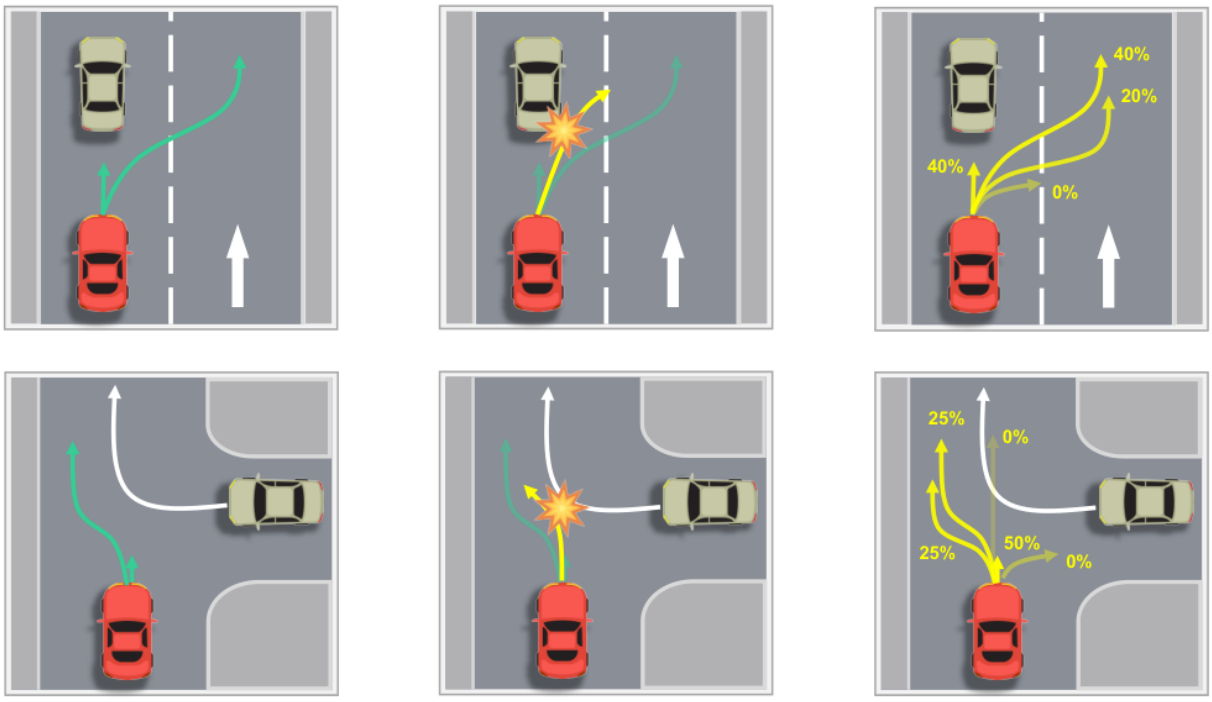
Abstract

大規模な運転デモから人間のような運転方針を学習することは有望であるが、計画の不確実性と非決定性のため、困難である。本研究では、不確実性問題に対処するため、確率的プランニングに基づくエンドツーエンドの運転モデルであるVADv2を提案する。VADv2 は、マルチビュー画像シーケンスをストリーミングで入力とし、センサデータを環境トークン埋め込みに変換し、行動の確率的分布を出力し、車両を制御するために 1 つの行動をサンプリングする。カメラセンサのみで、VADv2はCARLA Town05ベンチマークで最先端のクローズドループ性能を達成し、全ての既存手法を大幅に上回る。ルールベースのラッパーなしでも、完全にエンドツーエンドで安定的に動作する。閉ループデモは https://hgao-cv.github で公開されている。io/VADv2.

1. Introduction

エンドツーエンドの自律走行は、近年、重要かつ人気のある分野である。人間の運転デモの質量は簡単に入手できる。大規模なデモから人間のような運転方針を学習することは有望であると思われる。
しかし、プランニングの不確実性と非決定性により、運転デモから運転知識を抽出することは困難である。このような不確実性を実証するために、図1に2つのシナリオを示す。1）別の車両に追従する。人間のドライバーは、車線の追従や変更を追従させながら追い越すなど、多様な合理的な運転操作を行う。2）来るべき車両とのインタラクション人間のドライバーには2つの運転操作が可能です、を獲得するか、追い越すかである。統計学の観点からは、行動(タイミングと速度を含む)は非常に確率的であり、モデル化できない多くの潜在的な要因に影響される。
既存の学習ベースの計画手法[23, 19, 21, 40, 16, 54]は、決定論的なパラダイムに従って、行動を直接回帰する。回帰目標a^は、[23, 19, 21, 40]では未来の軌道、[16, 54]では制御信号(加速とステアリング)である。このようなパラダイムは、環境と行動の間に決定論的な関係が存在することを前提としているが、そうではない。人間の運転行動のばらつきが回帰対象の曖昧さを引き起こしている。特に、実行可能解空間が非凸の場合(図1参照)、決定論的モデリングは非凸の場合に対応できず、中間的な動作を出力する可能性があり、安全性の問題を引き起こす。また、このような決定論的回帰に基づくプランナは、学習データに最も現れる支配的な軌道を出力する傾向があり(停止や直進など)、望ましくない計画性能をもたらす。
本研究では、計画の不確実性に対処するために、確率的計画を提案する。我々の知る限り、VADv2は確率的モデリングを用いて連続計画行動空間を適合させた最初の研究であり、これは決定論的モデリングを用いて計画を行うこれまでの手法とは異なる。計画方針を環境条件付き非定常確率過程としてモデル化し、p(a|o)として定式化する。ここで、oは運転環境の履歴と現在の観測値、aは計画行動の候補である。決定論的モデリングと比較して、確率論的モデリングは、計画における不確実性を効果的に捉え、より正確で安全な計画性能を達成することができる。
計画行動空間は高次元の連続時空間空間である。行動空間から確率分布へのマッピングをモデル化するために、確率的な場関数に頼る。連続的な計画行動空間を直接フィッティングすることは不可能であるため、

⋆ Equal contribution
✉ Corresponding author: xgwang@hust.edu.cn
This work is still in process.



ドライビングデモンストレーション　　決定論的プランニング　　確率的プランニング

図1. 計画には不確実性が存在する。環境と行動の間には決定論的な関係は存在しない。決定論的計画は、特に実行可能解空間が非凸である場合、このような不確実性をモデル化できない。VADv2は確率的プランニングに基づき、大規模な運転デモから環境条件付き確率的行動分布を学習する。

計画行動空間を大きな計画語彙に離散化し、大量運転デモを用いて計画語彙に基づく計画行動の確率分布を学習する。離散化のために、走行デモのすべての軌跡を収集し、最も遠い軌跡サンプリングを採用して、計画語彙となるN個の代表的な軌跡を選択する。

確率的計画には他に2つの利点がある。まず、確率的プランニングは、各行動と環境との相関をモデル化する。目標計画行動に対してのみスパースな監視を提供する決定論的モデリングとは異なり、確率的計画は、正サンプルだけでなく、計画語彙のすべての候補に対しても監視を提供することができ、より豊かな監視情報をもたらす。また、確率的プランニングは推論段階で柔軟である。マルチモード計画の結果を出力し、ルールベースや最適化ベースの計画手法と組み合わせることが容易である。また、行動空間全体の分布をモデル化しているため、計画語彙に他の計画行動候補を柔軟に追加し、評価することができる。

確率的プランニングに基づき、我々はエンドツーエンドの運転モデルであるVADv2を発表する。

VADv2は、サラウンドビューの画像シーケンスをストリーミング方式で入力とし、センサデータをトークン埋め込みに変換し、行動の確率的分布を出力し、車両を制御するために1つの行動をサンプリングする。カメラセンサのみで、VADv2はCARLA Town05ベンチマークで最先端のクローズドループ性能を達成し、全ての既存手法を大幅に上回る。豊富なクローズドループデモは https://hgao-cv.github.io/VADv2 で紹介されている。VADv2はルールベースのラッパーなしでも、完全にエンドツーエンドで安定して動作する。

我々の貢献は以下のように要約される:

- プランニングの不確実性に対処するために、確率的プランニングを提案する。行動空間から確率分布に写像する確率場を設計し、大規模な運転デモから行動分布を学習する。
- 確率的プランニングに基づき、センサーデータを環境トークン埋め込みに変換し、行動の確率的分布を出力し、車両を制御するために1つの行動をサンプリングするエンドツーエンドの運転モデルであるVADv2を提案する。
- CARLAシミュレータでは、VADv2が最先端を達成している。

Town05ベンチマークにおけるクローズドループの性能。クローズドループのデモでは、エンドツーエンドで安定的に動作することがわかる。

2. Related Work

知覚。知覚は自律走行を実現するための最初のステップであり、運転シーンの統一的な表現は、下流のタスクに容易に統合するために有益である。鳥瞰図(BEV)表現は近年一般的な戦略となっており、効果的なシーン特徴の符号化とマルチモーダルデータフュージョンを可能にする。LSS[38]は、画像ピクセルの深度を明示的に予測することで、BEV変換への透視図を実現する先駆的な研究である。一方、BEVFormer [26, 52]は、空間的・時間的注意メカニズムを設計することで、明示的な奥行き予測を回避し、印象的な検出性能を達成している。その後の研究[25, 48]では、時間モデリングとBEV変換戦略を最適化することで、下流タスクの性能を継続的に向上させている。ベクトル化マッピングの観点から、HDMapNet [24]は、後処理により車線分割をベクトルマップに変換する。VectorMapNet [32]は、自己回帰的にベクトルマップ要素を予測する。MapTR [29, 30]は順列同値と階層マッチング戦略を導入し、マッピング性能を大幅に向上させる。LaneGAP [28]は、レーングラフのパスワイズモデリングを導入している。

モーション予測。モーション予測は、運転シーンにおける他の交通参加者の将来の軌跡を予測することを目的とし、自車両が情報に基づいた計画決定を行うのを支援する。従来の運動予測タスクは、過去の軌跡や高精細地図などの入力を利用して、将来の軌跡を予測する。しかし、近年のエンドツーエンドの動き予測手法[17, 53, 14, 22]の発展により、知覚と動き予測が共同で行われるようになった。シーン表現に関しては、ラスタライズされた画像表現を採用し、予測にCNNネットワークを採用した研究もある[3, 37]。他のアプローチでは、ベクトル化された表現を利用し、特徴抽出と動き予測にグラフニューラルネットワーク[27]やトランスフォーマーモデル[13, 33, 36]を採用している。いくつかの作品[17, 53]は、エージェントレベルの将来のウェイポイントではなく、将来のモーションを密な占有率とフローとして捉えている。いくつかの運動予測手法[14, 22]は、マルチモード軌道を回帰するためにガウス混合モデル(GMM)を採用している。不確実性をモデル化するための計画に適用できる。しかし、モードの数は限られている。

プランニング。学習ベースのプランニングは、そのデータ駆動型であり、データ量の増加に伴い素晴らしい性能を発揮するため、最近大きな可能性を示している。初期の試み[39, 8, 41]では、センサーデータを制御信号を予測するために直接使用する、完全にブラックボックス化されたスピリットを使用している。しかし、この戦略は解釈可能性に欠け、最適化が困難である。また、強化学習とプランニングを組み合わせた研究も数多く行われている[46, 5, 4]。閉ループシミュレーション環境における運転行動を自律的に探索することで、これらのアプローチは人間レベルの運転性能を達成、あるいは上回る。しかし、シミュレーションと現実のギャップを埋めるだけでなく、安全性の懸念に対処することは、強化学習戦略を実際の運転シナリオに適用する上で課題となる。模倣学習[2, 18, 19, 23]も研究の方向性の一つで、モデルは専門家の運転行動を学習することで、良好な計画性能を達成し、人間に近い運転スタイルを開発する。近年、知覚、運動予測、プランニングを一つのモデルに統合したエンドツーエンドの自律走行が登場し、その結果、完全にデータ駆動型のアプローチが実現し、有望な性能を発揮している。UniAD [19]は、複数の知覚タスクと予測タスクを巧みに統合し、計画性能を向上させる。VAD [23]は、密なマップの計画と除去のためのベクトル化されたシーン表現の可能性を探る。

自律走行における大規模言語モデル大規模言語モデル(LLM)が示す解釈可能性と論理的推論能力は、自律走行の分野に大いに役立つ。最近の研究では、LLMと自律走行の組み合わせが検討されている[7, 10, 12, 44, 51, 34, 31, 50, 49]。LLMを質問応答(QA)タスクを通してシーンの理解と評価を推進するために利用する研究がある。もう一つのアプローチは、LLMベースのシーン理解の上にプランニングを組み込むことで、さらに一歩進んでいる。例えば、DriveGPT4 [51]は、履歴ビデオやテキスト(質問や履歴制御信号のような追加情報を含む)などの入力を取得する。エンコード後、これらの入力はLLMに供給され、質問と制御信号に対する答えを予測する。一方、LanguageMPC [44]は、過去の真実の知覚結果とHDマップを言語記述の形で取り込む。次に、シーンを理解するために思考連鎖分析アプローチを利用し、LLMは最終的に予め定義されたセットから計画行動を予測する。各アクションは、実行のための特定の制御信号に対応する。VADv2はGPT[42, 43, 1, 47]からヒントを得て、不確実性問題に対処する。言語モデリングにも不確実性が存在する。特定の文脈が与えられたとき、次の単語は非決定論的で確率論的である。LLMは大規模なコーパスから次の単語の文脈条件付き確率分布を学習し、その分布から1つの単語をサンプリングする。LLMに触発され、VADv2は計画方針を環境条件付き非定常確率過程としてモデル化する。VADv2は行動空間を離散化して計画語彙を生成し、大規模な走行デモに基づいて確率分布を近似し、各時間ステップで分布から1つの行動をサンプリングして車両を制御する。

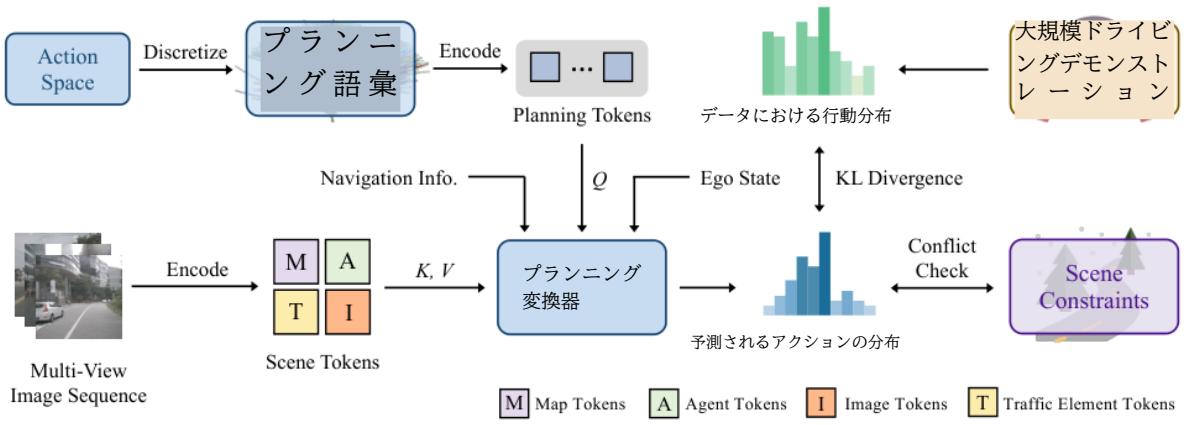


図2. VADv2の全体アーキテクチャ。VADv2は、マルチビュー画像シーケンスをストリーミングで入力とし、センサデータを環境トークン埋め込みに変換し、行動の確率分布を出力し、車両を制御するために1つの行動をサンプリングする。予測された分布を監視するために、大規模な運転デモとシーン制約が使用される。

3. Method

VADv2の全体的なフレームワークを図2に示す。VADv2は、マルチビュー画像シーケンスをストリーミングで入力とし、センサデータを環境トークン埋め込みに変換し、行動の確率分布を出力し、車両を制御するために1つの行動をサンプリングする。予測された分布を監視するために、大規模な運転デモとシーン制約が使用される。

3.1. Scene Encoder

画像の情報が疎で低レベルである。エンコーダを用いてセンサデータをインスタンスレベルのトークン埋め込みE_{env}に変換し、高レベル情報を明示的に抽出する。E_{env}には、マップトークン、エージェントトークン、トラフィック要素トークン、画像トークンの4種類のトークンが含まれる。VADv2は、地図トークンのグループ[30, 29, 28]を利用して、地図のベクトル表現(車線中心線、車線分割、道路境界、歩行者横断を含む)を予測する。また、VADv2はエージェントトークン群[22, 26]を用いて、他の交通参加者の運動情報(位置、向き、サイズ、速度、マルチモードの未来軌跡を含む)を予測する。交通要素も計画立案に重要な役割を果たす。VADv2 は、センサデータをトラフィック要素トークンに変換し、トラフィック要素の状態を予測する。CARLAでは、信号機と一時停止標識の2種類の信号機を考える。地図トークン、エージェントトークン、交通要素トークンは、対応する上位情報を明示的にエンコードするように、対応する監視信号で監視される。また、豊富な情報を含み、上記のインスタンスレベルのトークンと補完的な画像トークンをシーン表現としてプランニングを行う。また、ナビゲーション情報と自我の状態も埋め込み{E_{navi}, E_{state} }にエンコードされる。

MLP.

3.2. 確率的計画

プランニングの不確実性に対処するために、確率的プランニングを提案する。計画方針を環境条件付き非定常確率過程としてモデル化し、p(a|o)として定式化する。計画行動空間を大規模な走行デモに基づく確率分布として近似し、各時間ステップで分布から1つの行動をサンプリングして車両を制御する。

計画行動空間は高次元連続時空間空間$_A$ = {a|a$\in R^{2T}$ }である。連続的な計画行動空間を直接フィッティングすることは実行不可能であるため、計画行動空間を大きな計画語彙$_V$ = {a^i }^{N} に離散化する。具体的には、走行デモにおける全ての計画行動を収集し、最も遠い軌跡サンプリングを採用して、計画語彙となる代表的なN個の行動を選択する。V}の各軌跡は走行デモからサンプリングされるため、自車両の運動学的制約を自然に満たす。つまり、軌跡を制御信号(ステア、スロットル、ブレーキ)に変換した場合、制御信号の値は実行可能範囲を超えない。デフォルトでは、Nは4096に設定されている。

計画語彙の各アクションをウェイポイント列 a = (x_1, y_1, x_2, y_2, ・・・, x_T, y_T) として表現する。各ウェイポイントは将来のタイムスタンプに対応する。確率p(a)はaに関して連続であり、aのわずかな偏差に影響されない、すなわち、$\lim_{a\to 0}$ [p(a) − p(a + Δa)] = 0であると仮定する。5次元空間(x, y, z, θ, ϕ)上の連続輝度場をモデル化するNeRF [35]に触発されて、行動空間$_A$から確率分布{p(a)|a$\in_A$への連続マッピングをモデル化する確率場に頼る。

各行動(軌跡)を高次元計画トークン埋め込みE(a)に符号化し、カスケードTr環境情報E_{env}との相互作用のためのansformerデコーダを用い、ナビゲーション情報E_{navi}と自我状態E_{state}と組み合わせて確率を出力する、すなわち、

$$
\begin{aligned}
&p(a) = \mathrm{MLP}(\mathrm{Transformer}(E(a), E_{\mathrm{env}}) + E_{\mathrm{navi}} + E_{\mathrm{state}}),\\
&q = E(a), k = v = E_{\mathrm{env}},\\
&a = (x_1, y_1, x_2, y_2, ..., x_{\mathrm{T}}, y_{\mathrm{T}}),\\
&E(a) = \mathrm{Cat}\Big[\Gamma(x_1), \Gamma(y_1), \Gamma(x_2), \Gamma(y_2), ..., \Gamma(x_{\mathrm{T}}), \Gamma(y_{\mathrm{T}})\Big],\\
&\Gamma(pos) = \mathrm{Cat}\Big[\gamma(pos, 0), \gamma(pos, 1), ..., \gamma(pos, L-1)\Big],\\
&\gamma(pos, j) = \mathrm{Cat}\Big[\cos(pos/10000^{2\pi j/L}, \sin(pos/10000^{2\pi j/L}\Big].
\end{aligned}
\tag{1}
$$

～は$_R$の各座標を高次元の埋め込み空間$^{2L}_R$に写像する符号化関数であり、軌跡aの各座標値に個別に適用される。pos は位置を表す。これらの関数を用いて、連続入力座標を高次元空間にマッピングし、高周波フィールド関数をよりよく近似する。

3.3. Training

VADv2を3種類の監視、分布損失、競合損失、シーントークン損失で学習させる。

$$\mathcal{L} = \mathcal{L}_{\mathrm{distribution}} + \mathcal{L}_{\mathrm{conflict}} + \mathcal{L}_{\mathrm{token}}. \tag{2}$$

分布損失。大規模な運転デモから確率分布を学習する。KLダイバージェンスは、予測された分布とデータの分布との差を最小化するために使用される。

$$\mathcal{L}_{\mathrm{distribution}} = D_{\mathrm{KL}}(p_{\mathrm{data}} || p_{\mathrm{pred}}). \tag{3}$$

学習段階では、グランドトゥルースの軌跡を正サンプルとして計画語彙に追加する。その他の軌跡は負のサンプルとみなす。負の軌道に異なる損失重みを割り当てる。グランドトゥルースの軌跡に近い軌跡はペナルティが少ない。

コンフリクトロス。運転シーンの制約を利用して、モデルが運転に関する重要な事前知識を学習し、さらに予測分布を正則化するのを助ける。具体的には、計画語彙の中のある行動が他のエージェントの将来の動きや道路境界線と衝突する場合、その行動は負のサンプルとみなされ、この行動の確率を下げるために大きな損失重みを課す。

シーントークンの損失。マップトークン、エージェントトークン、トラフィック要素トークンは、対応する上位情報を明示的にエンコードするように、対応する監視信号で監視される。

地図トークンの損失はMapTRv2[30]と同じである。l_1損失は、予測マップ点とグランドトゥルースマップ点の間の回帰損失を計算するために採用される。焦点損失はマップ分類損失として使用される。

エージェントトークンの損失は、検出損失と動き予測損失から構成され、VAD[23]と同じである。l_1損失はエージェントの属性(位置、向き、大きさなど)を予測する回帰損失として、焦点損失はエージェントのクラスを予測する回帰損失として使用される。グランドトゥルースエージェントとマッチした各エージェントについて、K個の将来の軌道を予測し、最終的な変位誤差(minFDE)が最小となる軌道を代表的な予測値として使用する。次に、この代表的な軌跡とグランドトゥルースの軌跡の間のl_1損失を動き回帰損失として計算する。また、マルチモーダルな動き分類損失として、フォーカルロスを採用している。

交通要素トークンは、信号機トークンと停止標識トークンの2つの部分から構成される。一方では、信号機トークンをMLPに送信し、信号機の状態(黄、赤、緑)と信号機が自車両に影響を与えるかどうかを予測する。一方、ストップサイントークンもMLPに送られ、ストップサインエリアとエゴビークルの重なりを予測する。これらの予測を監督するために、フォーカルロスが使用される。

3.4. Inference

閉ループ推論では、分布から運転方針π_{model}を柔軟に得ることができる。直感的には、各時間ステップで最も確率の高い行動をサンプリングし、PIDコントローラを使用して、選択した軌道を制御信号(ステア、スロットル、ブレーキ)に変換する。

実世界のアプリケーションでは、確率分布をフルに活用するために、よりロバストな戦略がある。優れた実践例として、上位K個のアクションをプロポーザルとしてサンプリングし、プロポーザルのフィルタリングにはルールベースのラッパーを採用し、洗練には最適化ベースのポストソルバーを採用する。また、行動確率はエンドツーエンドモデルの信頼性を反映しており、従来のPnCと学習型PnCを切り替えるための判断条件とみなすことができる。

4. Experiments

4.1. 実験設定

VADv2の性能評価には、広く使われているCARLA[11]シミュレータを採用した。一般的な手法に従い、閉ループ評価にはTown05 LongとTown05 Shortベンチマークを使用する。具体的には、各ベンチマークにはあらかじめ定義された複数の走行ルートが含まれている。Town05 Longは10路線で構成され、各路線の長さは約1kmである。Town05 Shortは32のルートで構成され、各ルートの長さは70mである。

Method	Modality	Reference	Driving Score ↑	ルート完了 ↑ 上位	Infraction Score ↑
CILRS [9]	C	CVPR 19	7.8	10.3	0.75
LBC [6]	C	CoRL 20	12.3	31.9	0.66
Roach [54]	C	ICCV 21	41.6	96.4	0.43
Transfuser† [40]	C+L	TPAMI 22	31.0	47.5	0.77
ST-P3 [18]	C	ECCV 22	11.5	83.2	-
VAD [23]	C	ICCV 23	30.3	75.2	-
ThinkTwice [21]	C+L	CVPR 23	70.9	95.5	0.75
MILE [16]	C	NeurIPS 22	61.1	97.4	0.63
Interfuser [45]	C	CoRL 22	68.3	95.0	-
DriveAdapter+TCP [20]	C+L	ICCV 23	71.9	97.3	0.74
DriveMLM [49]	C+L	arXiv	76.1	98.1	0.78
VADv2	C	Ours	85.1	98.4	0.87

表1. Town05 Longベンチマークでのクローズドループ評価。

Method	Modality	Driving Score ↑	Route Completion ↑
CILRS [9]	C	7.5	13.4
LBC [6]	C	31.0	55.0
Transfuser [40]	C+L	54.5	78.4
ST-P3 [18]	C	55.1	86.7
VAD [23]	C	64.3	87.3
VADv2	C	89.7	93.0

表2. Town05 Shortベンチマークでのクローズドループ評価。

Town05 Longはモデルの総合的な能力を検証し、Town05 Shortは交差点前の車線変更など、特定のシナリオにおけるモデルの性能評価に重点を置いている。

Town03、Town04、Town06、Town07、Town10の運転ルートをランダムに生成し、CARLAの公式自律エージェントを用いて学習データを収集する。データは2Hzの周波数でサンプリングされ、学習用に約300万フレームを収集する。各フレームについて、6台のカメラによるサラウンドビュー画像、交通信号、他の交通参加者に関する情報、エゴ車両の状態情報を保存する。さらに、CARLAが提供するOpenStreetMap [15]フォーマットのマップを前処理することで、オンラインマッピングモジュールを学習するためのベクトル化マップを得る。なお、地図情報は学習時にグランドトゥルースとして提供されたものであり、VADv2はクローズドループ評価において高精細地図を利用していない。

4.2. Metrics

閉ループ評価には、CARLAの公式メトリクスを使用する。ルート完了は、エージェントが完了したルート距離の割合を示す。屈折スコアは、ルートに沿って起こる屈折の程度を示す。典型的な違反行為には、赤信号の走行、歩行者との衝突などが含まれる。各違反の種類に対応するペナルティ係数があり、違反が多いほど違反スコアは低くなる。

Driving Scoreは、Route CompletionとInfraction Scoreの積として機能し、評価の主要な指標となる。ベンチマーク評価では、ほとんどの作品がルールベースのラッパーを採用し、違反を減らす。他の手法と公平に比較するために、学習ベースのポリシーよりもルールベースのラッパーを採用する一般的な慣例に従う。

オープンループ評価では、L2距離と衝突率を採用し、学習されたポリシーがエキスパートデモにどの程度似ているかを示す。アブレーション実験では、オープンループメトリクスの計算が速く、安定していることを考慮し、オープンループメトリクスを評価に採用した。CARLAの公式自律エージェントを用いて、Town05 Longベンチマークの検証セットを生成し、オープンループ評価を行い、その結果を全検証サンプルの平均値とした。

4.3. 最先端手法との比較

Town05 Longベンチマークにおいて、VADv2は、Tab.1に示すように、Drive Score 85.1、Route Completion 98.4、Infraction Score 0.87を達成した。1. VADv2は、従来の最先端手法[49]と比較して、Drive Scoreを9.0大幅に改善しながら、より高いRoute Completionを達成している。なお、[49]がカメラとLiDARの両方を利用しているのに対し、VADv2はカメラのみを知覚入力として利用している。さらに、カメラのみに依存する従来の最良の方法[45]と比較して、VADv2はさらに大きな利点を示し、Drive Scoreが最大16.8と顕著に向上した。

Town05 Shortベンチマークに関するすべての公開研究の結果をTab. 2. Town05 Longベンチマークと比較して、Town05 Shortベンチマークは、混雑した交通流における車線変更や交差点前の車線変更など、特定の運転行動を実行するモデルの能力を評価することに重点を置いている。

ID	Dist. Loss	Conflict Loss	Agent Token	Map Token	Traf. Elem. Token	Image Token	L2 (m) ↓ 1s	2s	3s	Collision (%) ↓ 1s	2s	3s
1		✓	✓	✓	✓	✓	1.415	2.310	3.153	0.698	0.755	0.746
2	✓		✓	✓	✓	✓	0.086	0.173	0.291	0.0	0.012	0.039
3	✓	✓		✓	✓	✓	0.089	0.190	0.327	0.015	0.047	0.085
4	✓	✓	✓		✓	✓	0.086	0.191	0.332	0.005	0.034	0.070
5	✓	✓	✓	✓		✓	0.082	0.171	0.295	0.000	0.017	0.051
6	✓	✓	✓	✓	✓		0.083	0.170	0.293	0.000	0.010	0.039
7	✓	✓	✓	✓	✓	✓	0.082	0.169	0.290	0.000	0.010	0.039

表3. デザイン選択のためのアブレーション「Dist. Loss "は分布損失を表す。"Traf. Elem. Token "はトラフィック要素トークンを表す。

前の結果[23]と比較すると、VADv2はDrive Scoreを25.3、Route Completionを5.7と大幅に改善し、複雑な運転シナリオにおけるVADv2の総合的な運転能力を実証している。

4.4. Ablation Study

表 3はVADv2の主要モジュールのアブレーション実験である。このモデルは、分布損失(ID 1)によって提供される専門家の運転行動の監視なしには、計画精度の点で劣る。Conflict Lossは運転に関する重要な事前情報を提供するため、Conflict Loss(ID 2)がなければ、モデルの計画精度も影響を受ける。シーントークンは重要なシーン要素を高次元の特徴にエンコードし、計画トークンはシーントークンと相互作用して、走行シーンに関する動的情報と静的情報の両方を学習する。シーントークンのいずれかのタイプが欠落している場合、モデルの計画性能に影響します(ID 3-ID 6)。モデルが前述の設計をすべて組み込んだときに、最高の計画性能が達成される(ID 7)。

4.5. Visualization

図3は、VADv2の定性的な結果を示している。最初の画像は、異なる走行速度でVADv2によって予測されたマルチモーダルな計画軌道を示している。2番目の画像は、車線変更シナリオにおける前方クリーピングとマルチモーダル左折軌道の両方に対するVADv2の予測を示している。3番目の画像は、交差点における右車線変更シナリオを示しており、VADv2は、直進と右車線変更の両方について複数の軌道を予測する。最終画像は、ターゲットレーンに車両が存在する車線変更シナリオを示し、VADv2は複数の合理的な車線変更軌道を予測する。

5. Conclusion

本研究では、確率的プランニングに基づくエンドツーエンドの運転モデルであるVADv2を紹介する。

CARLAシミュレータにおいて、VADv2は安定して動作し、最先端の閉ループ性能を達成した。この確率的パラダイムの実現可能性は、主に検証されている。しかし、より複雑な実世界のシナリオにおける有効性はまだ未解明であり、今後の課題である。

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[2] Sergio Casas, Abbas Sadat, and Raquel Urtasun. Mp3: A unified model to map, perceive, predict and plan. In *CVPR*, 2021. 3

[3] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv preprint arXiv:1910.05449*, 2019. 3

[4] Raphael Chekroun, Marin Toromanoff, Sascha Hornauer, and Fabien Moutarde. Gri: General reinforced imitation and its application to vision-based autonomous driving. *arXiv preprint arXiv:2111.08575*, 2021. 3

[5] Dian Chen, Vladlen Koltun, and Philipp Krähenbühl. Learning to drive from a world on rails. In *ICCV*, 2021. 3

[6] Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. 2020. 6

[7] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. *arXiv preprint arXiv:2310.01957*, 2023. 3

[8] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *ICCV*, 2019. 3

[9] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. 2019. 6

[10] Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. *arXiv preprint arXiv:2309.05186*, 2023. 3

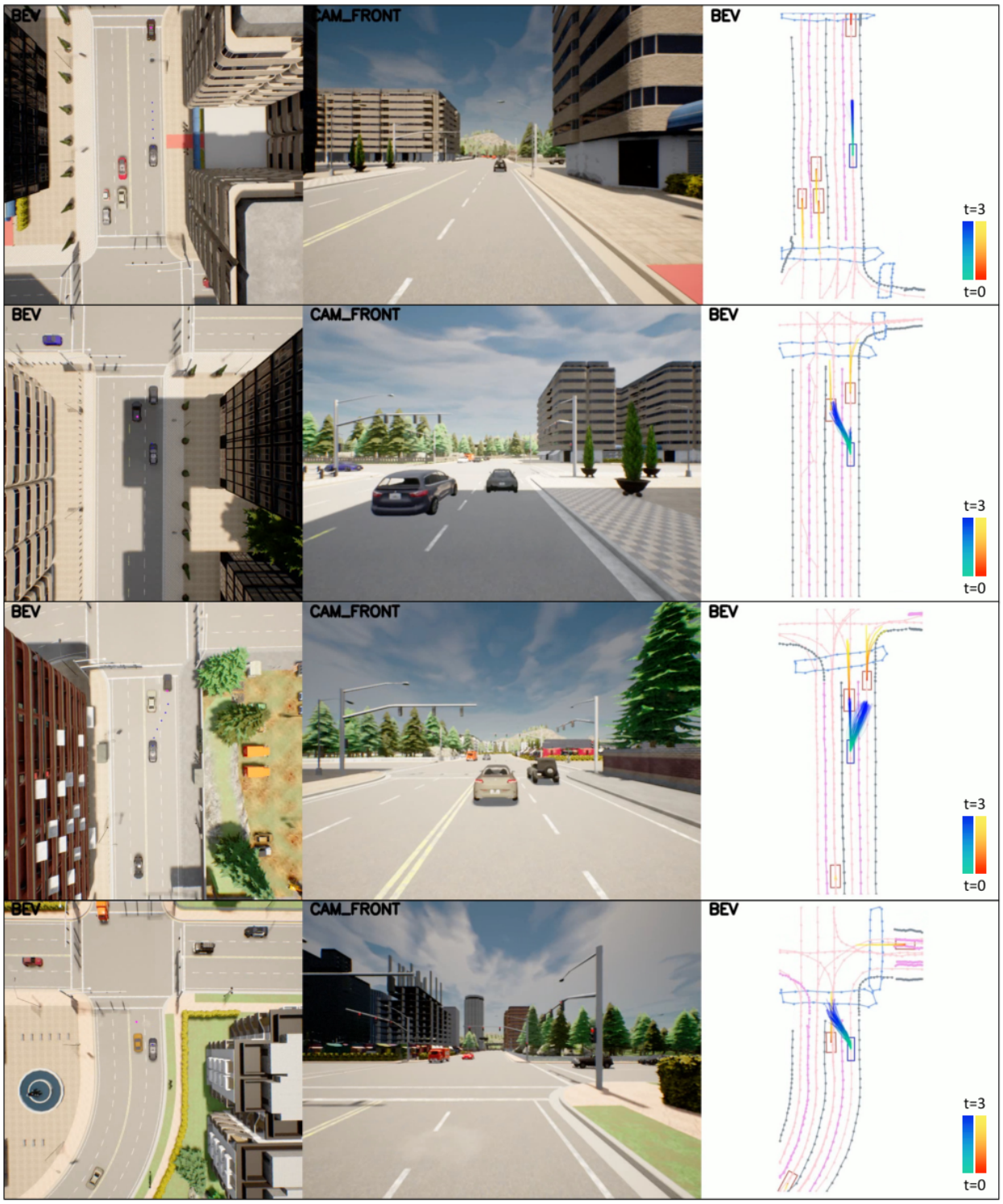


図3. VADv2の定性的結果。

[11] アレクセイ・ドソヴィツキー、ドイツ・ロス、フェリペ・コデヴィラ、アントニオ・ロペス、ヴラドレン・コルトゥン。カルラ:都市の開放的なドライブングシミュレータ。ロボット学習学会，ページ 1-16．PMLR，2017．5

[12] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. *arXiv preprint arXiv:2307.07162*, 2023. 3

[13] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *CVPR*, 2020. 3

[14] Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. *arXiv preprint arXiv:2208.01582*, 2022. 3

[15] Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive computing*, 2008. 6

[16] Anthony Hu, Gianluca Corrado, Nicolas Griffiths, Zak Murez, Corina Gurau, Hudson Yeo, Alex Kendall, Roberto Cipolla, and Jamie Shotton. Model-based imitation learning for urban driving. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 6

[17] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *ICCV*, 2021. 3

[18] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *ECCV*, 2022. 3, 6

[19] 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. *CVPR2023*, 2022. 1, 3

[20] Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. 2023. 6

[21] Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In *CVPR*, 2023. 1, 6

[22] Bo Jiang, Shaoyu Chen, Xinggang Wang, Bencheng Liao, Tianheng Cheng, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, and Chang Huang. Perceive, interact, predict: Learning dynamic and static clues for end-to-end motion prediction. *arXiv preprint arXiv:2212.02181*, 2022. 3, 4

[23] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. *ICCV*, 2023. 1, 3, 5, 6, 7

[24] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *ICRA*, 2022. 3

[25] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022. 3

[26] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: 時空間変換器を用いたマルチカメラ画像からの鳥瞰表現の学習。arXivプレプリントarXiv:2203.17270, 2022. 3, 4

[27] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *ECCV*, 2020. 3

[28] Bencheng Liao, Shaoyu Chen, Bo Jiang, Tianheng Cheng, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Lane graph as path: Continuity-preserving path-wise modeling for online lane graph construction. *arXiv preprint arXiv:2303.08815*, 2023. 3, 4

[29] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. *arXiv preprint arXiv:2208.14437*, 2022. 3, 4

[30] Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Maptrv2: An end-to-end framework for online vectorized hd map construction. *arXiv preprint arXiv:2308.05736*, 2023. 3, 4, 5

[31] Jiaqi Liu, Peng Hang, Jianqiang Wang, Jian Sun, et al. Mtd-gpt: A multi-task decision-making gpt model for autonomous driving at unsignalized intersections. *arXiv preprint arXiv:2307.16118*, 2023. 3

[32] Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. *arXiv preprint arXiv:2206.08920*, 2022. 3

[33] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. In *CVPR*, 2021. 3

[34] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 3

[35] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *ECCV*, 2020. 4

[36] Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, et al. Scene transformer: A unified architecture for predicting multiple agent trajectories. *arXiv preprint arXiv:2106.08417*, 2021. 3

[37] Tung Phan-Minh, Elena Corina Grigore, Freddy A Boulton, Oscar Beijbom, and Eric M Wolff. Covernet: Multimodal behavior prediction using trajectory sets. In *CVPR*, 2020. 3

[38] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, 2020. 3

[39] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *NeurIPS*, 1988. 3

[40] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. 2021. 1, 6

[41] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *CVPR*, 2021. 3

[42] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 3

[43] 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. 3

[44] Hao Sha, Yao Mu, Yuxuan Jiang, Li Chen, Chenfeng Xu, Ping Luo, Shengbo Eben Li, Masayoshi Tomizuka, Wei Zhan, and Mingyu Ding. Languagempc: Large language models as decision makers for autonomous driving. *arXiv preprint arXiv:2310.03026*, 2023. 3

[45] Hao Shao, Letian Wang, Ruobing Chen, Hongsheng Li, and Yu Liu. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In *Conference on Robot Learning*, pages 726–737. PMLR, 2023. 6

[46] Marin Toromanoff, Emilie Wirbel, and Fabien Moutarde. End-to-end model-free reinforcement learning for urban driving using implicit affordances. In *CVPR*, 2020. 3

[47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[48] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. *arXiv preprint arXiv:2303.11926*, 2023. 3

[49] Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023. 3, 6

[50] Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models. *arXiv preprint arXiv:2309.16292*, 2023. 3

[51] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth KY Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*, 2023. 3

[52] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. *arXiv preprint arXiv:2211.10439*, 2022. 3

[53] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022. 3

[54] Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. End-to-end urban driving by imitating a reinforcement learning coach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 6