潜在世界モデルによるエンド・ツー・エンドの自律走行の強化

Yingyan Li[1,2] **Lue Fan**[1,2] **Jiawei He**[1,2] **Yuqi Wang**[1,2] **Yuntao Chen**[3]

Zhaoxiang Zhang[1,2,3✉] **Tieniu Tan**[1,2]

[1] Institute of Automation, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] CAIR, HKISI, CAS
Email: liyingyan2021@ia.ac.cn
Code: https://github.com/BraveGroup/LAW

Abstract

エンド・ツー・エンドの自律走行が広く注目されている。現在のエンドツーエンドのアプローチは、シーン表現の学習を支援するために、検出、追跡、マップセグメンテーションなどの知覚タスクからの監視に大きく依存している。しかし、これらの方法は広範なアノテーションを必要とし、データのスケーラビリティを妨げている。この課題を解決するために、我々は、高価なラベルを必要とせずにエンドツーエンドの運転を強化する新しい自己教師付き手法を提案する。具体的には、我々のフレームワークLAWは、予測された自我行動と現在のフレームの潜在特徴に基づいて、将来の潜在特徴を予測するためにLATent Worldモデルを使用する。予測された潜在特徴量は、将来実際に観測された特徴量によって教師される。このスーパービジョンは、潜在特徴学習と行動予測を共同で最適化することで、運転性能を大幅に向上させる。その結果、我々のアプローチは、コストのかかるアノテーションを行うことなく、オープンループとクローズドループの両方のベンチマークで最先端の性能を達成した。

1 Introduction

エンドツーエンドの自律走行[15, 22, 30, 40, 14]は、従来の方法と比較して潜在的な利点があることが認識されつつある。従来のプランナーは、元のセンサーデータにアクセスできない。これは情報損失とエラーの蓄積につながる[15, 22]。一方、エンドツーエンドのプランナーは、センサーデータを処理して、計画決定を直接出力するため、さらなる探求のための有望な分野であることが示された。

ほとんどのエンドツーエンドの自律走行手法[15, 22, 14, 30]は、エンドツーエンドで動作するものの、検出、追跡、地図分割などの様々な補助タスクを活用している。これらの補助的なタスクは、モデルがより良いシーン表現を学習するのに役立つ。しかし、手作業によるアノテーションが多く、非常に高価であり、データのスケーラビリティに限界がある。一方、いくつかのエンドツーエンド手法[35, 4, 40]は、知覚タスクを採用せず、記録された運転ビデオと軌跡のみから学習する。これらのアプローチは、利用可能な大量のデータを活用することができ、有望な方向性である。しかし、軌跡からの限られたガイダンスしか使用しないため、ネットワークが効果的なシーン表現を学習し、最適な運転性能を達成することは困難である。

この問題に対処するため、図1に示すように、自己教師あり学習によってエンドツーエンドの運転を強化する。従来の画像処理における自己教師付き手法[10, 6]は、一般的に静的な単一フレーム画像に集中している。

Preprint. Under review.

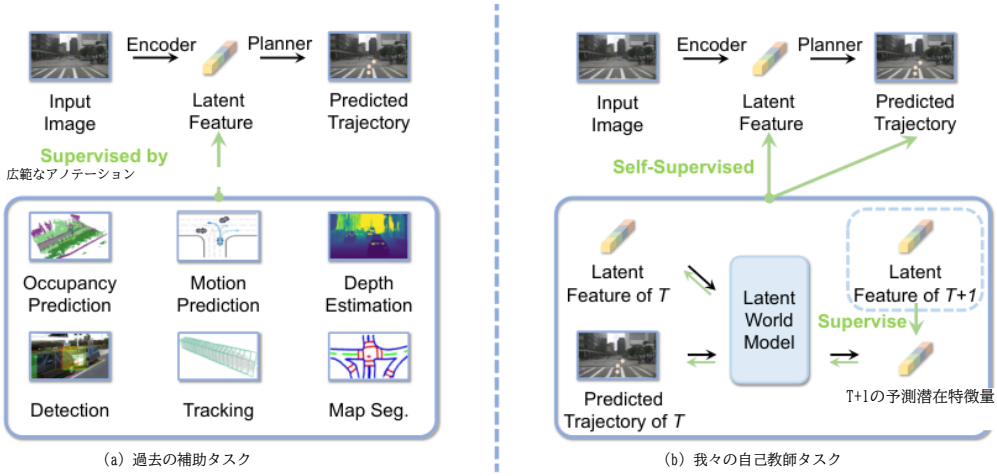


図1:これまでの補助タスクと我々の潜在予測タスクの比較。(a)の先行研究が広範な注釈を持つ補助的な知覚タスクに依存しているのに対して、我々は(b)の潜在世界モデルを通してエンドツーエンドの運転モデルを強化することを目的とする。学習時には、未来フレームから潜在特徴を取得し、現在のフレームの潜在特徴と予測軌跡を共同で監視する。セグメンテーション

しかし、自律走行には動的な一連の入力が含まれるため、時間データを効果的に利用することが不可欠である。運転における重要なスキルは、現在の周囲環境に基づいて将来の状態を予測することである。これに触発され、我々は潜在特徴の予測を目的とした自己教師タスクを提案する。具体的には、現在の状態とエゴの行動に基づいて将来の状態を予測する潜在世界モデルを開発し、状態はネットワーク内の潜在的なシーン特徴として表現される。学習時には、潜在世界モデルから予測される潜在特徴を監督するために、未来フレームの潜在特徴を抽出する。その結果、現在のフレームの潜在特徴学習と軌跡予測を共同で最適化する。

さらに、ビューワイズ潜在特徴を抽出するためのシンプルかつ強力なプランナを確立し、提案する潜在世界モデルのテストベッドとして機能する。これまでの手法とは異なり、このプランナはアドホックなモジュールや知覚に関連する分岐を組み込んでいないため、潜在世界モデルの内部構造を理解することがより容易である。このプランナーと潜在世界モデルを考えると、副産物がある。潜在世界モデルは未来のビューの潜在特徴を予測することができるので、未来フレーム内のいくつかのビューの特徴抽出プロセスをスキップし、これらのビューの予測された未来を置換として使用することができる。特定のビューの特徴抽出をスキップすることで、パイプライン全体の効率を向上させる。どのビューを置換すべきかを決定するために、ビュー選択戦略を提案する。ビュー潜在置換と組み合わせることで、この戦略は最小限の性能損失でパイプライン全体を大幅に高速化する。
要約すると、我々の主な貢献は以下の通りである:

- 我々は、エンドツーエンドの自律走行フレームワークの学習を強化する、自己教師付き学習のためのLATent Worldモデルを提案する。
- 潜在世界モデルに基づき、我々はさらにビュー選択戦略を提案し、最小限の性能損失を発生させながらパイプラインを大幅に高速化する。
- 我々のフレームワークLAWは、オープンループとクローズドループの両方のベンチマークにおいて、手動アノテーションなしで最先端の結果を達成した。

2 Related Works

2.1 エンド・ツー・エンドの自律走行

エンドツーエンドの自律走行手法[15, 22, 31, 35]を、従来の知覚タスクを実行するかどうかによって、明示的手法と暗黙的手法の2つに分類する。

明示的なエンドツーエンド手法[2, 30, 19, 34]は、検出[24, 17]、追跡[45, 37]、マップ分割[15, 22]、占有予測[38, 18]など、複数の知覚タスクを同時に実行する。先駆的な研究として、P3 [32]は、運動計画プロセスのコスト要因として、微分可能な意味的占有表現を採用している。これに続いて、ST-P3 [14]は、知覚、予測、計画タスクのためのより代表的な特徴を同時に生成する空間-時間特徴学習アプローチを導入している。そして、多くの研究[15, 30, 20, 19]は、BEV特徴マップに基づく検出とBEVマップ分割タスクの実行に焦点を当てている。代表的なものとして、UniAD [15]は、トラッキングや動作予測を含む複数のモジュールを統合し、ゴール駆動型プランニングをサポートする。VAD [22]は、計画目的のためにベクトル化されたシーン表現を探求する。
暗黙のエンドツーエンド手法[35, 4, 43, 40]は、多数の知覚注釈を利用することを避けるため、有望な方向性を示す。初期の暗黙のエンドツーエンド手法[43, 35]は、主に強化学習に依存していた。例えば、MaRLn [35]は暗黙のアフォーダンスに基づく強化学習アルゴリズムを設計し、LBC [4]は特権(真実の知覚)情報を用いて強化学習エキスパートを学習した。強化学習エキスパートによって生成された軌跡データを用いて、TCP [40]は軌跡ウェイポイント分岐と直接制御分岐を組み合わせ、良好な性能を達成した。しかし、暗黙のエンドツーエンドの手法は、しばしば不十分なシーン表現能力に悩まされる。我々の研究は、潜在的な予測によってこの問題に対処することを目的としている。

2.2 自律走行における世界モデル

自律走行における既存の世界モデルは、画像ベースの世界モデルと占有率ベースの世界モデルの2種類に分類できる。画像ベースの世界モデル[11, 39, 12]は、生成的アプローチによって自律走行データセットを充実させることを目的としている。GAIA-1[12]は、ビデオ、テキスト、アクションの入力を利用して、現実的な運転シナリオを作成する生成世界モデルである。MILE [11]は、3Dジオメトリを帰納的バイアスとして活用することで、都市走行映像を生成する。Drive-WM[39]は、拡散モデルを利用して、これらの予測画像に基づいて将来の画像と計画を予測する。Copilot4D [42]は、VQVAE [36]を用いてセンサー観測をトークン化し、離散拡散によって未来を予測する。もう一つのカテゴリーは、占有率に基づく世界モデル[44, 29]である。OccWorld[44]とDriveWorld[29]は、占有率を予測するためにワールドモデルを使用する。逆に、我々の提案する潜在世界モデルは、手動による注釈を必要としない。

3 Preliminary

エンドツーエンド自律走行エンドツーエンド自律走行のタスクでは、目的はウェイポイントの形で自車両の将来の軌道を推定することである。形式的には、$_t$ = {I_t^1, I_t^2,.. I_t^N } を時間ステップtで撮影されたN枚の周囲マルチビュー画像の集合とする。モデルはウェイポイントのシーケンスW_t = {w_t^1, w_t^2,.. ここで、各ウェイポイント w_t^i = (x it, y_t^i) は、時間ステップ t + i における自車両の予測 BEV 位置を表す。Mは、モデルが予測しようとする自車両の将来の位置の数を表す。
世界モデル自律走行タスクにおいて、世界モデルは現在の状態と行動に基づいて将来の状態を予測することを目的とする。具体的には、時間ステップtで現在のフレームから抽出された特徴量をF^$_t$とすると、W_t = {w_t^1, w_t^2,.. w_t^M }はプランナによる計画されたウェイポイントのシーケンスを表し、ワールドモデルはF^$_t$とW_tを用いて未来フレームの特徴F^$_{t+1}$を予測する。

4 Method

全体的な方法論は3つの部分に分けられる。まず、4.1節で潜在[1]を抽出するための強力で一般的なエンドツーエンドプランナを開発する。次に、エンドツーエンドプランナーに基づき、4.2節で潜在を予測するワールドモデルを紹介する。最後に、予測された潜在は重要でない潜在を代替することができるので、Sec.4.3ではビュー選択アプローチを提案する。

4.1 潜在抽出を用いたエンドツーエンドプランナー

効果的な潜在特徴を抽出するために、一般的で強力なエンドツーエンドのプランナを導入する。最初に、Nビューの画像は、それぞれの特徴表現を抽出するために、画像バックボーンを通して処理される。

[1]The terms "latent" and "latent feature" convey the same meaning.

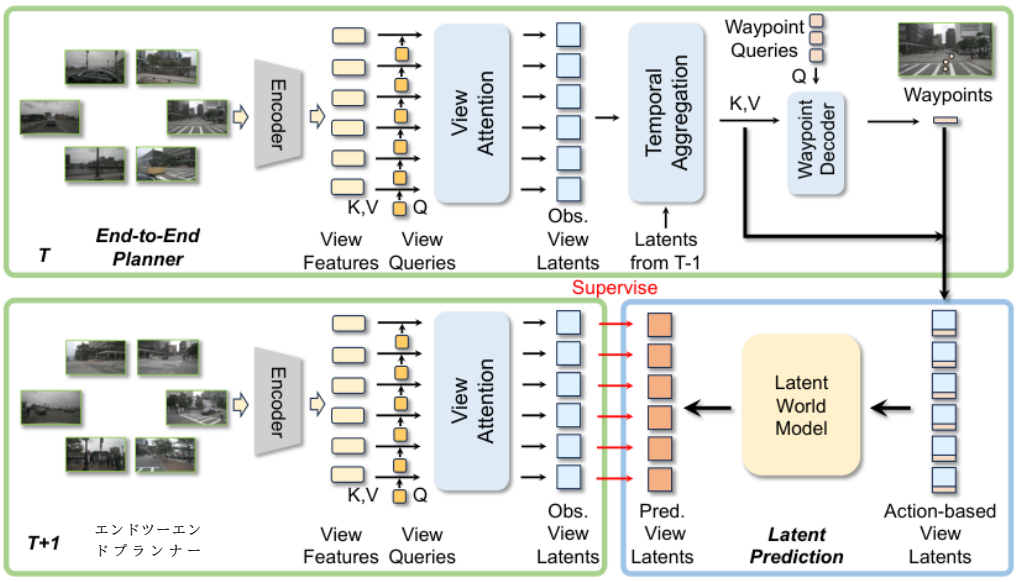


図2:全体的なフレームワーク具体的には、ビューの潜在能力を抽出し、ウェイポイントを予測するためのエンドツーエンドのドライビングフレームワークを開発する。次に、潜在世界モデルによって次のフレームのビュー潜在を予測する。予測されたビュー潜在は、次のフレームの観測されたビュー潜在によって監督される。観測値：観測値、予測値：予測値。

PETR[25]に従い、これらの画像特徴に対して3次元位置埋め込みを生成する。これらの位置埋め込みは、画像特徴量と統合され、各ビューを一意に識別する。濃縮された画像特徴を$F = \{f^1, f^2, .. f^N\}$とする。

次に、Fを観測されたビュー潜在Vに圧縮するために、ビュー注意メカニズムを採用する。ここでは、このビュー潜在を、後述する他のものと区別するために、「観測された」という用語を使用する。具体的には、N個のビューに対して、学習可能なビュークエリ$Q_{view} = \{q^1_{view}, q^2_{view}, .. q^N_{view}\}$である。各ビュークエリq iviewは、対応する画像特徴f^iとクロスアテンションを受け、N個の観測ビュー潜在$V = \{v^1, v^2, .. v^N\}$となる。

$$\mathbf{v}^i = \text{CrossAttention}(\mathbf{q}^i_{view}, \mathbf{f}^i, \mathbf{f}^i). \tag{1}$$

f^iはクロスアテンションのキーと値として機能する。次に、観測されたビュー潜在量に対して時間的集約を行う。観測されたビュー潜在量Vは、前のフレームから生成された履歴ビュー潜在量Hによって強調される(詳細は4.2節で述べる)。このようにして

$$\mathbf{E} = \mathbf{V} + \mathbf{H}, \tag{2}$$

ここで、Eはenhanced view latentと名付けられる。Eが与えられたとき、ウェイポイントをデコードするためのウェイポイントデコーダを開発する。具体的には、M個のウェイポイントクエリ$Q_{wp} = \{q^1_{wp}, q^2_{wp}, .. q^M_{wp}\}$、ここで各クエリは学習可能な埋め込みである。これらのウェイポイントクエリは、クロスアテンションメカニズムを通じてEと相互作用する。更新されたウェイポイントクエリはMLPヘッドに渡され、ウェイポイント$W = \{w^1, w^2, .. w^M\}$を出力する：

$$\mathbf{w}^j = \text{MLP}(\text{CrossAttention}(\mathbf{q}^j_{wp}, \mathbf{E}, \mathbf{E})). \tag{3}$$

学習中、L1損失を用いて、予測されたウェイポイントとグランドトゥルースのウェイポイントとの間の不一致を次のように測定する：

$$\mathcal{L}_{\text{waypoint}} = \sum_{j=1}^{M} \|\mathbf{w}^j_t - \mathbf{w}^{j,\text{GT}}_t\|_1, \tag{4}$$

提案するエンドツーエンドプランナは、潜在的な世界をシンプルかつ効果的に抽出し、潜在的な世界モデルの良いテストベッドとして機能する。

4.2 潜在予測のための世界モデル

本節では、潜在世界モデルを利用して、未来フレームのビュー潜在を予測する。まず始めに、拡張ビュー潜在量E_tと予測ウェイポイントW_tに基づいて、アクションベースのビュー潜在量を生成する。具体的には、E_t = {e_t^1, e_t^2 ,.. e_t^N }とすると、W_t = {w_t^1, w_t^2,.. w_t^M }を一次元ベクトル$w_{et} \in R^{2M}$に変換する。次に、特徴チャネル次元に沿ってe itとw_{et}を連結する。連結されたベクトルは MLP によって変換され、e it の特徴チャネル次元に一致する a it を形成する。形式的には、i番目のビューのアクションベースビュー潜在は、次のように示される：

$$\mathbf{a}_t^i = \mathrm{MLP}([\mathbf{e}_t^i, \widetilde{\mathbf{w}}_t]), \tag{5}$$

ここで、[-, -]は連結演算を表す。全体の行動ベースのビュー潜在は、A_t = {a_t^1, a_t^2,.. a_t^N }である。続いて、A_tが与えられると、潜在世界モデルによってフレームt + 1の予測ビュー潜在P_{t+1}が得られる：

$$\mathbf{P}_{t+1} = \mathrm{LatentWorldModel}(\mathbf{A}_t). \tag{6}$$

潜在世界モデルのネットワークアーキテクチャは、2つのブロックからなる変換デコーダである。各ブロックには自己注意とFFNモジュールが含まれる。自己アテンションはビュー次元で実行される。学習中、エンドツーエンドのプランナを用いて、フレームt + 1の観測されたビュー潜在V_{t+1}を抽出する。V_{t+1}はL2損失関数を用いてP_{t+1}の監視の役割を果たす：

$$\mathcal{L}_{\text{latent}} = \sum_{i=1}^{N} \|\mathbf{p}_{t+1}^i - \mathbf{v}_{t+1}^i\|_2, \tag{7}$$

where $\mathbf{P}_{t+1} = \{\mathbf{p}_{t+1}^1, \ldots, \mathbf{p}_{t+1}^N\}$ and $\mathbf{V}_{t+1} = \{\mathbf{v}_{t+1}^1, \ldots, \mathbf{v}_{t+1}^N\}$.

また、A_tが与えられると、時間情報を履歴ビュー潜在H_{t+1}にエンコードする。H_{t+1}は、式(2)により、観測されたビュー潜在V_{t+1}を強調するために用いられる。具体的には、ビュー次元のA_tに対して自己アテンションを行い、以下のようになる。

$$\mathbf{H}_{t+1} = \mathrm{SelfAttention}(\mathbf{A}_t). \tag{8}$$

H_{t+1}とP_{t+1}は異なる関数を持つ。H_{t+1}は時間情報を残差として符号化することを目的とし、P_{t+1}は未来フレームのビュー潜在を予測するように設計されている。また、P_{t+1}は、将来フレームの観測されたビュー潜在の良い代替となり、潜在代替を用いたビュー選択の概念を提案するきっかけとなる。

4.3 潜在置換によるビュー選択

世界モデルによって予測される有効なビュー潜在能力を利用したビュー選択アプローチを提案する。マルチビュー動画を入力として、このアプローチは動的にいくつかの有益なビューを選択し、特徴を抽出する。他のビューは処理されず、対応するビュー潜在量はワールドモデルから予測されるビュー潜在量で代用される。図3に示すように、本節は3つの構成要素から構成されている。まず、いくつかの潜在的なビュー選択戦略が与えられると、選択報酬予測コンポーネントはこれらの戦略の報酬を予測し、最も高い報酬を持つ戦略を選択する。次に、選択されたビューを持つプランナーは、選択されたビューが与えられたときに軌跡を予測する。学習中、各選択戦略に報酬ラベルを割り当てる選択報酬ラベリングモジュールを提案する。

選択報酬予測図3(a)、(b)に示すように、各選択戦略に関連する報酬を推定するために設計された報酬予測モジュールを導入する。報酬は、各戦略を用いて得られた計画結果の有効性を定量的に反映する。詳細には、K個の選択クエリを定義する。これらのK個の選択クエリは、K個の潜在的な選択戦略に対応する。各選択クエリは学習可能な埋め込みである。各戦略は、処理のために特定のビューを選択し、残りは破棄する。次に、クエリと世界モデルによって予測されたビュー潜在P_{t+1}との間のクロスアテンションを実行することによって、選択クエリを更新する。更新された選択クエリは、報酬を予測するためにMLPヘッドに供給される。これらの報酬が与えられたとき、我々は最も高い予測報酬を持つ戦略を選択する。戦略はフレームtで選択され、この戦略によって選択されたビューは、フレームt + 1で選択されたビューを持つプランナーの入力として機能する。

このプランナは、図3(c)が示すように、選択されたビューを入力としてウェイポイントを生成する。これは4.1節のプランナーと同じ重みを共有する。

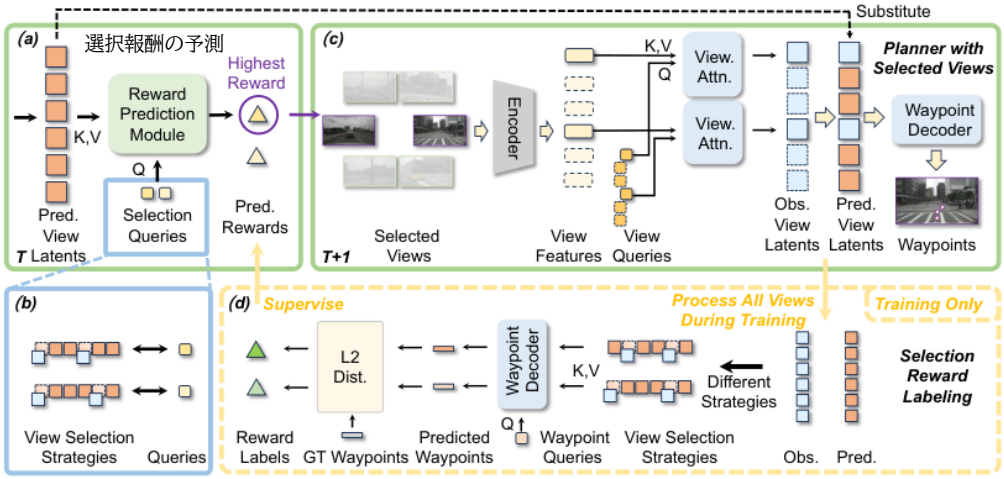


図3:ビュー選択によるエンドツーエンドドライビングのパイプライン。予測されたビュー潜在が与えられると、(a)は(b)で定義された選択クエリに対する報酬を予測する。次に、ビューを選択するために、(c)のプランナの入力となる、最も報酬の良い選択戦略を採用する。学習中、各選択戦略にラベルを付けるために、(d)の選択報酬ラベリングモジュールを提案する。Pred: 予測、Obs: 観測。L2距離: L2距離。

具体的には、Nをビューの総数とする。$S \subseteq \{1,2,..N\}$を、選択されたビューのインデックスの集合と定義する。$v_i \in E_{t+1}$と$\hat{v}_j \in P_{t+1}$が与えられたとき、合成ビュー潜在量$v_{combined}$は次のように定式化される:

$$\mathbf{v}_{combined} = \bigoplus_{i \in S} \mathbf{v}_i \oplus \bigoplus_{j \in \bar{S}} \hat{\mathbf{v}}_j, \tag{9}$$

ここで^{L}はビュー次元の潜在能力に対する連結演算を表す。次に、結合されたビュー潜在は、軌跡を予測するために、セクション4.1の同じパイプラインに渡される。選択ビューを持つプランナーに基づいて、選択戦略にラベルを付ける選択報酬ラベリングモジュールを提案する。

選択報酬ラベリング図3(d)に示すように、報酬ラベリングのアプローチを導入する。具体的には、k番目の戦略について、対応する選択されたビューが、選択されたビューを持つプランナーに供給され、ウェイポイント$\hat{w}_k$を予測する。k番目の戦略の報酬ラベルd_k^*は、予測されたウェイポイント$\hat{w}_k$とグランドトゥルースのウェイポイント$\hat{w}^{GT}$の間のL2距離として、次のように定義される:

$$\hat{d}_k = -\|\hat{\mathbf{w}}_k - \hat{\mathbf{w}}^{GT}\|_2. \tag{10}$$

The larger $\hat{d}_k$ is, the closer $\hat{\mathbf{w}}_k$ are to the ground truth waypoints $\hat{\mathbf{w}}^{GT}$. During training, we use the L1 loss to learn the rewards, formulated as $\mathcal{L}_{\text{reward}} = \sum_{k=1}^{K} ||d_k - \hat{d}_k||_1$,

要約すると、我々のフレームワークの総損失は

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{waypoint}} + \mathcal{L}_{\text{latent}} + \mathcal{L}_{\text{reward}}, \tag{11}$$

ここで、L_{reward}はビュー選択アプローチを使用するかどうかに依存するオプションの損失である。各損失の重みについては、実装の詳細で説明する。

5 Experiments

5.1 Setup

オープンループベンチマーク オープンループベンチマークは、専門家であるドライバーのビデオストリームと、それに対応する自車両の軌跡を記録する。我々は、1,000の運転シーンからなるnuScenesデータセット[1]を用いて実験を行う。

表1:オープンループnuScenes [1]ベンチマークにおける最先端手法との比較。ST-P3、VAD、LAWのFPSはNVIDIA Geforce RTX 3090 GPUでテスト。UniADのFPSはNVIDIA Tesla A100 GPUでテスト。‡ : LiDARベースの手法。過去のエゴの状態情報を使用しない。

Method	L2 (m) ↓				Collision (%) ↓				Latency (ms)	FPS
	1s	2s	3s	Avg.	1s	2s	3s	Avg.		
NMP‡ [41]	-	-	2.31	-	-	-	1.92	-	-	-
SA-NMP‡ [41]	-	-	2.05	-	-	-	1.59	-	-	-
FF‡ [13]	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43	-	-
EO‡ [23]	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33	-	-
ST-P3 [14]	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71	628.3	1.6
UniAD [15]	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31	555.6	1.8
VAD [22]	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22	224.3	4.5
LAW	0.26	0.57	1.01	0.61	0.14	0.21	0.54	0.30	51.2	19.5

先行研究[14, 16, 22]と同様に、計画性能を総合的に評価するために、変位誤差(DE)と衝突率(CR)を採用する。変位誤差は、予測された軌道とGT軌道の間のL2距離を測定する。衝突率は、予測された軌道をたどったときに、他の物体と衝突する割合を定量化する。

閉ループ評価(Closed-loop Benchmark) 閉ループ評価は、運転動作に基づいてセンサー入力を常に更新するため、自律走行に不可欠である。学習データセットは、[40, 20]に従い、教師モデルRoach [43]を用いてCARLA [9]シミュレータ(バージョン0.9.10.1)から収集され、189Kフレームとなる。クローズドループ走行性能の評価には、広く使われているTown05 Longベンチマーク[20, 33, 11]を用いる。公式メトリクスを使用する: ルート完了(RC)は、自律エージェントが完了したルートの割合を表す。走行距離スコア(IS)は、走行距離の違反だけでなく、走行距離の違反の数を定量化する。Infraction Scoreが高いほど、安全な運転方法の遵守度が高いことを示す。ドライビングスコア(DS)は、総合的なパフォーマンスを評価するために使用される主要な指標である。ルート完了と違反スコアの積として計算される。

表2:CARLAにおけるクローズドループTown05 Longベンチマークの性能。エキスパート:特権的なエキスパートの走行軌跡から模倣学習を行う。Seg.: セマンティックセグメンテーション。Map: BEVマップセグメンテーション。Dep: 奥行き推定。Det.: 3次元物体検出。潜在予測:我々の提案する自己教師タスク。

Method	Supervision	DS↑	RC↑	IS↑
CILRS [7]	Expert	7.8±0.3	10.3±0.0	0.75±0.05
LBC [5]	Expert	12.3±2.0	31.9±2.2	0.66±0.02
Transfuser [30]	Expert, Dep., Seg., Map., Det.	31.0±3.6	47.5±5.3	0.77±0.04
Roach [43]	Expert	41.6±1.8	96.4±2.1	0.43±0.03
LAV [3]	Expert, Seg., Map., Det.	46.5±2.3	69.8±2.3	0.73±0.02
TCP [40]	Expert	57.2±1.5	80.4±1.5	0.73±0.02
MILE [11]	Expert, Map., Det.	61.1±3.2	97.4±0.8	0.63±0.03
ThinkTwice [21]	Expert, Dep., Seg., Det.	65.0±1.7	95.5±2.0	0.69±0.05
DriveAdapter [20]	Expert, Map., Det.	65.9±-	94.4±-	0.72±-
Interfuser [33]	Expert, Map., Det.	68.3±1.9	95.0±2.9	-
LAW	Expert, *Latent Prediction*	70.1±2.6	97.8±0.9	0.72±0.03

実装の詳細 LAWのデフォルトの設定には、指定がない限りビューの選択は含まれていない。オープンループベンチマークでは、Swin-Transformer-Tiny [26] (Swin-T)をバックボーンとして使用する。入力画像は800×320にリサイズされる。初期学習率を5e-5として、コサインアニーリング[27]学習率スケジュールを採用する。AdamW[28]オプティマイザを0.01の重み減衰で利用し、バッチサイズ8で8個のRTX 3090 GPUで12エポック学習する。ウェイポイント損失と潜在予測損失の重みは1.0に設定される。選択されたビューを持つプランナについては、LAWに基づく報酬損失で微調整を行う。初期学習率を5e-6に設定し、さらに6エポック学習する。報酬損失の重みは1.0に設定される。クローズドループベンチマークの場合。また、公平な比較のために、[40]に従いResNet-34をバックボーンとして使用する。

表3:潜在予測に関するアブレーション研究。潜在世界モデルは、ビュー潜在と予測された軌跡の2種類の入力を受け取る。入力なしとは、ワールドモデルを利用しないことを意味する。Agg.: 集約。Pred.: 予測。Traj: 軌跡

Temporal Agg.	ワールドモデルの入力		L2 (m) ↓				Collision (%) ↓			
	View Latent	Pred. Traj.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
-	-	-	0.44	0.95	1.65	1.01	0.27	0.57	1.32	0.72
✓	-	-	0.32	0.67	1.14	0.71	0.20	0.30	0.73	0.41
✓	✓	-	0.30	0.64	1.12	0.68	0.18	0.27	0.66	0.37
✓	✓	✓	0.26	0.57	1.01	0.61	0.14	0.21	0.54	0.30

TCPヘッド[40]を[20]に準じて使用する。入力画像のサイズは900×256である。オプティマイザはAdamである。学習率は1e-4、重み減衰は1e-7とする。バッチサイズ128で60エポック学習させる。30エポック後に学習率が2分の1に減少する。

5.2 最先端手法との比較

オープンループベンチマークでは、nuScenesデータセットにおいて、LAWをUniAD [15]、VAD [22]などの最先端手法と比較する。結果を表1にまとめる。LAWは、1s、2s、3sの予測ホライズンにおける平均L2変位誤差の点で、UniADとVADを上回る。さらに、本手法は30.9msのレイテンシで顕著なリアルタイム性能を達成し、本手法の効率性を浮き彫りにした。クローズドループベンチマークでは、表2に示すように、我々の提案手法は全ての既存手法を凌駕している。注目すべきは、我々のアプローチは、奥行き推定、セマンティックセグメンテーション、マップセグメンテーションからの広範な監視を組み込んだThinkTwice [21]やDriveAdapter [20]などの以前の主要な手法を凌駕していることである。

5.3 アブレーション研究

潜在予測(Latent Prediction) このアブレーション研究では、潜在予測の有効性を調査する。結果を表3に示す。当初は、予測された軌跡成分を省略したビューラマンのみを使用した。tentを入力とするワールドモデルは、表3の3行目に示すように、このアプローチは潜在予測なしのモデルと比較して、わずかな性能向上をもたらす。予測された軌跡を入力の一部として含めると(表3の4行目)、性能が大幅に向上する。将来の潜在能力を正確に予測するためには、運転行動を取り入れる必要があり、潜在能力世界モデルを使用することの合理性が強調されることを示している。さらに、表4に示すように、閉ループ設定における潜在予測に関するアブレーション研究を提供する。

表4:Town05 Longベンチマークにおける潜在予測に関するアブレーション研究。

Latent Prediction	DS↑	RC↑	IS↑
×	67.9±2.1	98.6±0.8	0.68±0.02
✓	70.1±2.6	97.8±0.9	0.72±0.03

注目すべきは、Infraction Scoreの大幅な改善が観察されたことである。これは、将来のシナリオを予測する能力が、潜在的な衝突を効果的に緩和するのに役立つことを示している。

潜在世界モデルのネットワークアーキテクチャ潜在世界モデルのネットワークアーキテクチャの影響を検証するために、表5に示すような実験を行う。まず、線形射影として表現される単層ニューラルネットワークは、世界モデルの機能を果たすのに十分ではなく、その結果、性能が著しく低下することが明らかである。2層MLPは性能の大幅な向上を示している。しかし、異なるビューからの潜在的な間の相互作用を促進する機能を欠いている。そこで、デフォルトのネットワークアーキテクチャとして、テストしたアーキテクチャの中で最も良い結果を得たトランスフォーマーデコーダを使用する。このことは、特定のビューに対して、隣接する複数のビューからの情報を組み込むことで、将来の潜在的な予測を強化できることを示唆している。

潜在的世界モデルの時間地平この実験では、世界モデルは3つの異なる将来の時間地平で潜在的特徴を予測する: 0.5秒、1.5秒、3.0秒である。これは、nuScenesデータセットでは0.5秒ごとにキーフレームが発生することを考えると、現在のフレームから1番目、3番目、6番目の未来のフレームに相当する。

表5:潜在世界モデルの異なるネットワークアーキテクチャに関するアブレーション研究。線形射影とは、単層ネットワークを意味する。Arch: アーキテクチャ

World Model Arch.	L2 (m) ↓				Collision (%) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
Linear Projection	0.31	0.65	1.14	0.70	0.26	0.34	0.66	0.42
Two-layer MLP	0.27	0.58	1.07	0.64	0.17	0.23	0.59	0.33
変換デコーダ	0.26	0.57	1.01	0.61	0.14	0.21	0.54	0.30

表6に示すように、1.5秒のホライズンにおいて、モデルが最高の性能を達成していることがわかる。その理由は以下の通りである。0.5秒間隔は通常、最小限の変更でシーンを提示し、特徴学習を改善するための動的コンテンツが不十分である。一方、3.0秒間隔では、予測タスクの複雑さが増し、より良い特徴学習の妨げとなる。この結論は、マスク比率が過度に低くても高くてもネットワークの能力に悪影響を与えるというMAE [10]の観察と一致する。

表6:潜在予測のための異なる時間軸に関するアブレーション研究。世界モデルは、現在から様々な将来の時間軸における潜在能力を予測することができる。

Time Horizon	L2 (m) ↓				Collision (%) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
0.5s	0.26	0.57	1.01	0.61	0.14	0.21	0.54	0.30
1.5s	0.26	0.54	0.93	0.58	0.14	0.17	0.45	0.25
3.0s	0.28	0.59	1.01	0.63	0.13	0.20	0.48	0.27

ビュー選択我々のビュー選択アプローチの有効性を排除するために、表7に示す実験を行う。ビュー選択モジュールでモデルを訓練し、いくつかの戦略でテストする: 1) 前面ビューとランダムビュー、2) 前面ビューとビュー選択モジュールによって選択されたビュー、3) 前面ビューと4.3節の報酬ラベルに基づいて選択されたビュー。この報酬ラベルはGT軌道の助けを借りて生成され、この実験が上限となる。4) 6つのビューすべて。フロントビューを固定する理由については付録A.1で説明する。その結果、我々のビュー選択モジュールによる選択は、ランダム選択を大幅に上回り、GTによる上限設定に密接に近づいていることが実証された。

表7:ビュー選択アプローチに関するアブレーション研究。選択報酬予測モジュールで学習したモデルを推論に使用する。GTビュー: 予測された報酬ではなく、報酬ラベルを用いたビュー選択戦略を採用する。

Selected views	L2 (m) ↓				Collision (%) ↓				Latency (ms) ↓
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	
Front + a random view	0.36	0.73	1.23	0.77	0.16	0.27	0.78	0.40	30.9
Front + predicted view	0.30	0.64	1.10	0.68	0.16	0.25	0.72	0.38	30.9
Front + GT view	0.28	0.56	0.97	0.60	0.15	0.22	0.61	0.33	30.9
Six views	0.26	0.57	1.01	0.61	0.14	0.21	0.54	0.30	51.2

6 結論と限界

結論として、本論文では潜在世界モデルを用いた新しい自己教師付きアプローチを紹介する。このアプローチは、コストのかかるアノテーションを必要としない、エンドツーエンドの自律走行システムにおけるシーン表現の学習を強化する。我々の手法は、現在のベンチマークで有望な結果を示しているが、利用されるデータ量が限られているため、制約がある。今後の課題として、より大規模で多様なデータセットに適用し、本アプローチのスケーラビリティを向上させることを目指す。大規模データを活用し、潜在世界モデルを事前学習に採用する予定である。

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[2] Sergio Casas, Abbas Sadat, and Raquel Urtasun. Mp3: A unified model to map, perceive, predict and plan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14403–14412, 2021.

[3] Dian Chen and Philipp Krähenbühl. Learning from all vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17222–17231, 2022.

[4] Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Conference on Robot Learning*, pages 66–75. PMLR, 2020.

[5] ディアン・チェン、ブレイディ・ズー、ヴラドレン・コルトゥン、フィリップ・クレーヘンビュール。不正行為による学習。CoRL，pages 66-75. PMLR，2020.

[6] ティン・チェン、サイモン・コーンブリス、モハマド・ノロウジ、ジェフリー・ヒントン。視覚表現の対比学習のためのシンプルなフレームワーク。機械学習国際会議、1597-1607 ページ。PMLR，2020.

[7] フェリペ・コデヴィラ、エデル・サンタナ、アントニオ・M・ロペス、アドリアン・ガイドン。自律走行のための行動クローニングの限界を探る。ICCV，2019.

[8] MMDetection3D 貢献者。MMDetection3D: OpenMMLab 一般的な3Dオブジェクト検出のための次世代プラットフォーム。https://github.com/オープンmmlab/mmdetection3d，2020.

[9] 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。Carla: オープンな都市型ドライビングシミュレータ。CoRL，2017.

[10] 何凱明、陳新雷、謝彩寧、李陽浩、ピョートル・ドラー、ロス・ガーシック。マスクオートエンコーダはスケーラブルな視覚学習器である。CVPR，16000-16009 ページ，2022.

[11]アンソニー・フー、ジャンルーカ・コラード、ニコラ・グリフィス、ザック・ムレス、コリーナ・グルー、ハドソン・ヨー、アレックス・ケンドール、ロベルト・チポラ、ジェイミー・ショットン。都市走行のためのモデルベース模倣学習。NeurIPS，2022.

[12] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

[13] Peiyun Hu, Aaron Huang, John Dolan, David Held, and Deva Ramanan. Safe local motion planning with self-supervised freespace forecasting. In *CVPR*, 2021.

[14] 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.

[15] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Goal-oriented autonomous driving. *arXiv preprint arXiv:2212.10156*, 2022.

[16] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *CVPR*, 2023.

[17] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

[18] 黄元輝、鄭文照、張雲鵬、周杰倫、呂吉文。視覚に基づく3次元意味的占有予測のための3視点ビュー。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9223-9232，2023.

[19] Bernhard Jaeger, Kashyap Chitta, and Andreas Geiger. Hidden biases of end-to-end driving models. In *ICCV*, 2023.

[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. In *ICCV*, 2023.

[21] ジアオソン、ウー・ペンハオ、チェン・リー、謝江偉、何コングイ、ヤン・ジュンチ、リ・ホンヤン。運転前に2回考える:エンドツーエンドの自律走行のためのスケーラブルなデコーダに向けて。2023年CVPR

[22] 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. In *ICCV*, 2023.

[23] Tarasha Khurana, Peiyun Hu, Achal Dave, Jason Ziglar, David Held, and Deva Ramanan. Differentiable raycasting for self-supervised occupancy forecasting. In *ECCV*, 2022.

[24] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022.

[25] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022.

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[27] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

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

[29] Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, et al. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. *arXiv preprint arXiv:2405.04390*, 2024.

[30] アディティヤ・プラカシュ、カシャップ・チッタ、アンドレアス・ガイガー。エンドツーエンドの自律走行のためのマルチモーダル融合変換器。CVPR, 2021.

[31] Katrin Renz, Kashyap Chitta, Otniel-Bogdan Mercea, A Koepke, Zeynep Akata, and Andreas Geiger. Plant: Explainable planning transformers via object-level representations. *arXiv preprint arXiv:2210.14222*, 2022.

[32] Abbas Sadat, Sergio Casas, Mengye Ren, Xinyu Wu, Pranaab Dhawan, and Raquel Urtasun. Perceive, predict, and plan: Safe motion planning through interpretable semantic representations. In *ECCV*, 2020.

[33] 趙浩、王レティアン、陳汝平、李洪生、劉宇。解釈可能なセンサーフュージョントランスを用いた安全性向上型自律走行。CoRL, 2022.

[34] Hao Shao, Letian Wang, Ruobing Chen, Steven L Waslander, Hongsheng Li, and Yu Liu. ReasonNet: End-to-End Driving with Temporal and Global Reasoning. In *CVPR*, pages 13723–13733, 2023.

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

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

[37] Qitai Wang, Yuntao Chen, Ziqi Pang, Naiyan Wang, and Zhaoxiang Zhang. Immortal tracker: Tracklet never dies. *arXiv preprint arXiv:2111.13672*, 2021.

[38] 王玉基、陳雲涛、遼興宇、ファン・ルー、張兆祥。Panoocc: カメラベースの3Dパノプティックセグメンテーションのための統一された占有率表現。arXivプレプリント arXiv:2306.10013, 2023.

[39] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. *arXiv preprint arXiv:2311.17918*, 2023.

[40] Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: a simple yet strong baseline. *NeurIPS*, 2022.

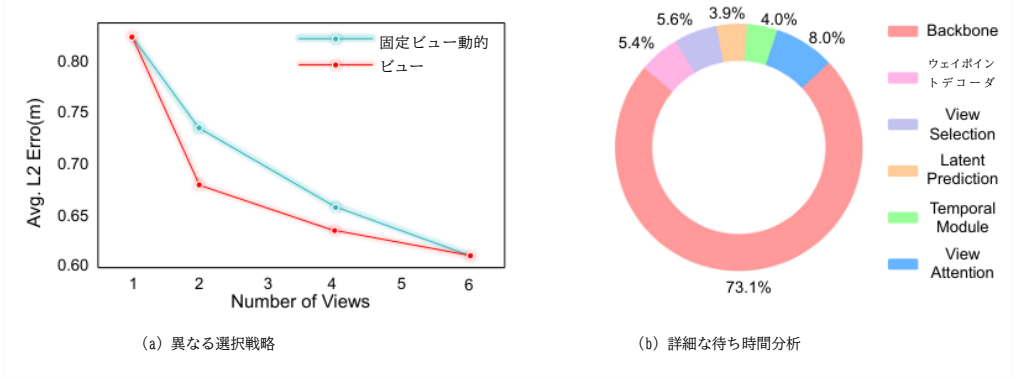
[41] Wenyuan Zeng, Wenjie Luo, Simon Suo, Abbas Sadat, Bin Yang, Sergio Casas, and Raquel Urtasun. End-to-end interpretable neural motion planner. In *CVPR*, 2019.

[42] 張倫軍、熊玉文、楊澤、セルジオ・カサス、胡瑞、ウルタスン・ラクエル。離散拡散による自律走行のための教師なし世界モデルの学習。arXivプレプリント arXiv:2311.01017, 2023.

[43] 張哲軍、アレクサンダー・リニガー、戴登信、フィッシャー・ユー、リュック・ヴァン・グール。強化学習コーチの模倣によるエンドツーエンドの都市走行。ICCV, 2021 にて。

[44] 鄭文照、陳偉良、黄元輝、張博瑞、段岳基、呂吉文。オクワールド: 自律走行のための3次元占有世界モデルの学習。arXivプレプリント arXiv:2311.16038, 2023.

[45] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European conference on computer vision*, pages 474–490. Springer, 2020.



（a）異なる選択戦略　　（b）詳細な待ち時間分析

図4:（a）異なる選択戦略間の比較。（b）詳細な待ち時間分析。

A Appendix

A.1 ビュー選択のさらなる分析

我々のビュー選択アプローチでは、常に正面ビューを使用し、他の5つのビューから1つの追加ビューを動的に選択する。その理由は、以下のような実験によってもたらされたものである。

フロントビューの固定は、常にフロントカメラビューを入力ビューの1つとして選択する実験的な正当性を示す。具体的には、報酬損失のない訓練されたエンドツーエンドプランナが与えられたとき、それに対して2つのビュー選択戦略を行う。最初の戦略は常にフロントカメラを使用し、残りの5つのカメラから1つのランダムなビューを選択する。第二の戦略は、ランダムに2台のカメラを選択する。結果を表8に示す。この実験から、正面ビューを固定し、追加ビューをランダムに選択する戦略は、2つのビューをランダムに選択する戦略よりも優れていることがわかる。この優位性は、特に前方走行シナリオにおいて、正面から見た方が計画タスクにとってより重要な情報を提供できることに起因している。

表8:フロントビューの重要性

Selected Views	L2 (m) ↓				Collision (%) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
Two random views	0.39	0.79	1.32	0.83	0.15	0.28	0.91	0.45
Front + a random view	0.36	0.74	1.24	0.78	0.16	0.28	0.82	0.42

異なる選択戦略により、固定カメラの数を減らすことで、計算負荷を軽減することもできる。したがって、ビュー選択アプローチが、より少ない固定カメラ数を利用した構成と比較してどうなのかを問うのは自然なことである。これを調べるために、図4(a)に示すような実験を行った。実験の具体的な設定は以下の通りである。ビュー数を1、2、4、6とし、4つの実験グループを設定した。1つのビューのみを使用する場合は、常にフロントカメラを使用する。2つのビューについて、固定ビューモデルはフロントカメラとバックカメラを使用して学習とテストを行い、ダイナミックビューモデル(すなわち、我々の方法)はフロントカメラを固定し、残りの5つのカメラから最も情報量の多いビューを選択する。4つのビューに対して、固定ビューモデルはフロントカメラ、フロント左カメラ、フロント右カメラ、バックカメラを使用し、ダイナミックビューモデルはフロントカメラを固定し、残りの5つからさらに3つのカメラを選択する。最後に、6つのビューについて、利用可能なすべてのカメラを使用する。この結果は、我々の動的選択法が、同じビュー数の固定ビュー設定を一貫して上回ることを示している。これは、我々の手法が複数の選択肢から有益なビューを選択できることを示している。

遅延内訳 Selected Views を用いたプランナーは、32.4 FPS という素晴らしい速度を達成した。ここでは、各モジュールの詳細なレイテンシを示す。NVIDIA Geforce RTX 3090 GPUでバッチサイズ1でテストする。コードはmmdetection3d [8]に基づいている。我々のモデルにおける各モジュールに関連する具体的な待ち時間は、図4(b)に詳述されている。図に示すように、バックボーンはモデルのレイテンシの大部分を占めている。

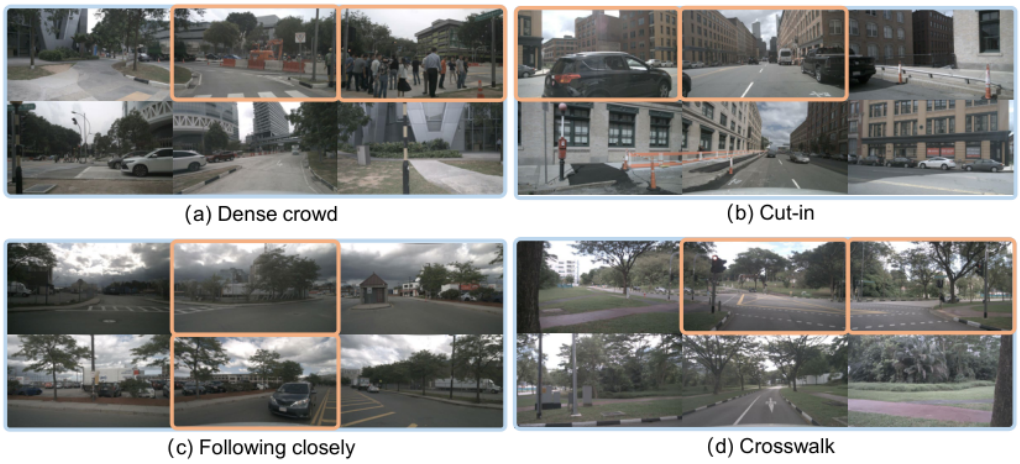


図5:ビュー選択の可視化。選択されたビューの概要をオレンジ色のボックスで示す。

入力ビューの数を減らすと、バックボーンのコストが直線的に減少する。バックボーンを通過するビューは2つしかないため、我々のビュー選択法は効率を大幅に向上させる。

ビュー選択の可視化 Fig.5 に示すように、4 つの典型的なケースの可視化分析を示す。これらの可視化から、我々は2つの重要な洞察を導き出した: 1)我々の手法は、視覚的に顕著なオブジェクトを持つビューを好む。(a)、(b)、(c)のケースで示されるように、モデルは、人、切断する可能性のある車両、または後端衝突のリスクのある車両のグループを特徴とするビューを選択する傾向がある。このような嗜好性は、このような物体が運転行動に大きな影響を与えるために生じる。報酬学習により、我々のモデルは、これらの顕著なオブジェクトによって提供される手がかりによって、ある瞬間に最も重要なビューを特定することができる。2)我々のモデルは、特定のシナリオに関する事前知識を学習する。人間のドライバーは、横断歩道で突然現れる歩行者の予想など、世界に関する予備知識を持っており、より遅い運転が必要である。我々のビュー潜在再構成モジュールは、同様の事前分布を学習することもできる。例えば、(d)の場合、再構成モジュールによって支援されるビュー選択モデルは、合理的に横断歩道に焦点を当てる。