時空間シーングラフを用いたGraph Attention Networksによる案内文生成の高精度化と視覚的説明の実現

鈴木 颯斗† 下村 晃太† 平川 翼† 山下 隆義† 藤吉 弘亘†
†中部大学
E-mail: hayato@mprg.cs.chubu.ac.jp

1 はじめに

自動運転技術や高度運転支援システムの発展に伴い，車両が運転手に対して適切な情報を提供することの重要性が高まっている．一般的なナビゲーションシステムは，地図情報に基づいた経路案内や，定型文による音声案内が主流である．しかし，これらは運転手の直感的な理解を必ずしも促すものではなく，特に走行環境の複雑さや動的な変化に対応した案内には限界がある．

このような課題に対し，Human-like Guidance に関する研究が注目されている．これは，車両周辺の状況に基づき，人間のように自然な表現で案内を行うことで，運転手の適切な判断を支援する技術である．従来手法としては，静的な地図情報を基にしたランドマーク選択の最適化手法 [1, 2] があり，交差点の形状や周辺建物情報を用いた案内を実現している．しかし，地図情報に依存するため，データ更新の頻度やランドマーク選定の柔軟性に課題があり，動的な環境変化や運転手の認知負荷への対応にも限界がある．このような背景から，車両の視界情報を活用して動的に環境理解を行う手法が考えられる．画像認識技術により走行環境を適切に把握することで，より柔軟で表現力の高い案内文の生成が期待される．

本研究では，Human-like Guidance の実現に向け，視界情報を基にした環境認識と自然言語生成を統合することで，運転手が直感的に理解可能な案内文を生成することを目標とする．走行シーンにおける案内の判断対象となる情報は多岐にわたり，特に時間的な変化を伴う環境認識では，情報量の増加が案内文の生成の精度や安定性に影響を及ぼす可能性がある．先行研究 [3] では，走行シーンの動画像から得られるオブジェクトの空間的・時間的関係をシーングラフとして表現することが有効であることが示されている．しかし，一般的に参照フレーム数が増えるほど判断精度が向上すると考えられる一方で，入力フレーム数の増加が性能に与える影響については十分に議論されていない．加えて，生成された案内文に対する視覚的な説明が提供されておらず，モデルの判断根拠が不明瞭である点も課題である．そこで本研究では，動画像からシンプルな構造のシーングラフを構築し，Graph Attention Networks (GAT)[4] をベースとしたテキスト生成モデルを提案する．推論時に得られる Attention 情報を可視化することで，生成文に対する視覚的説明を実現し，判断根拠を明確にすることでモデルの信頼性の向上を図る．評価実験では，従来の動画像処理手法との比較を行い，提案手法の有効性を検証する．また，異なる入力フレーム数に基づいたモデルの出力精度を比較することで，フレーム数の増加が生成性能に与える影響を検証する．

2 関連研究

本研究では，車両の視界情報を基に環境を理解し，適切な案内文を生成することを目的としている．この目的に関連する技術として，視覚情報を自然言語に変換する Image/Video Captioning や，オブジェクト間の関係性を表現する Graph Neural Network が挙げられる．本章ではこれらの関連研究について述べる．

2.1 Image/Video Captioning

Image Captioning とは，画像とテキストを関連付け，画像に対する自然言語の説明文を生成するタスクである．Image Captioing は主に，画像の特徴抽出を行う機構と，得られた画像特徴量から言語モデルを用いた文章生成を行う機構で構成される．従来では，show and tell[5] を始めとした CNN による特徴抽出と RNN ベースの言語モデルを組み合わせたアプローチが提案されている．Transformer[6] の登場以降，Self-Attention 機構を活用したアプローチが主流となっている．Transformer は，自然言語処理において優れた性能を発揮しており，その柔軟性の高さから様々なタスクに適していることが示されている．例として，言語モデルを RNN から Transformer に置き換えた手法 [7] が提案されている．また，画像認識の分野で Transformer を適用した Vision Transformer[8] は，パッチ分割した画像をトークンとして処理することで，大域的な特徴を捉えることが可能となった．これにより，CPTR[9] のような Transformer のみで End-to-End でキャプション生成を行う手法が提案されており，従来の Image Captioning 手法と比較し

て大幅に精度が向上した．

Video Captioning は，Image Captioning の枠組みを動画データに拡張したタスクであり，時間情報を考慮する点が大きな特徴である．動画データの特徴抽出を行うモデルとして，従来では，3DCNN[10] をベースとした手法が主流となっている．このタスクにおいても，Transformer を応用することが考えられており，動画全体の時間的関係を Self-Attention 機構で学習するアプローチが提案されている．SwinBERT[11] は，Transformer をベースとしてモデル全体を設計しており，柔軟な説明性を持つキャプションの生成を可能とした．

2.2 Graph Neural Network

Graph Neural Network[12] は，グラフ構造データの解析に適したニューラルネットワークであり，ノード間の関係やグラフの構造を考慮した情報伝播を行う．代表的な手法である Graph Convolutional Networks (GCN)[13] は，畳み込み演算をグラフに適用し，ノード分類やグラフ分類などで優れた性能を示した．GCN を拡張した Spatial Temporal GCN[14] は，時系列情報を持つグラフを学習することが可能となった．GAT[4] は，各ノードが異なる重みで隣接ノードの情報を集約できるようにするために，Attention 機構を導入している．これにより，ノード間の重要度を学習し，情報の伝播をより効果的に行うことが可能となった．

シーングラフを活用した手法 [15] では，画像内のオブジェクトをノードとして捉え，位置関係や意味的な繋がりをグラフ構造として表現している．Graph R-CNN[16] では，物体検出モデルが出力したオブジェクト間の関係性をシーングラフとして構築することで，画像の構造的理解が向上することを示した．

3 提案手法

本研究では，走行シーンの動画像から得られるオブジェクトの時空間的関係を時空間シーングラフとして構築し，GAT を Graph Enocoder として用いた Graph-to-Text モデルにより，運転手が直感的に理解しやすい案内文を生成する手法を提案する．シーングラフの構築には，詳細なオブジェクトクラスの検出が可能な物体検出モデルとトラッキング手法を組み合わせることで，空間および時系列の関係性を反映した単純なグラフ構造を生成する．さらに，交差点における進行方向などの特徴をグラフに統合することで，重要なオブジェクトへの着目を促す特徴量を導入する．また，Graph Encoder から得られる Attention 情報を可視化することで，案内文生成における判断根拠を視覚的に説明可能とし，モデルの解釈性と信頼性の向上を図る．提案手法の全体構成を図 1 に示す．本章では，提案手法の各構成要素について詳細に述べる．

3.1 マルチオブジェクトトラッキング

動画像中のオブジェクトを対象としたシーングラフ構築においては，オブジェクトの検出手法として YOLO-World[17] を用いる．YOLO-World はオープン語彙検出に対応しており，プロンプトで指定したオブジェクトを zero-shot で検出可能である．従来の物体検出モデルでは，事前定義されたカテゴリに依存するため，検出オブジェクトをノードとする際に色情報などの特徴を考慮できない．先行研究では，ノードに周辺領域の画像特徴量を含める手法を提案しているが，グラフ構造が複雑化する問題がある．一方で，YOLO-World は，“blue car” や “stop sign” といったより詳細なクラスを定義することができるため，それらを用いることでグラフ特徴の単純化が可能となる．

ただし，YOLO-World は静止画像における物体検出を対象としている．時系列を考慮したシーングラフを構築する場合，フレーム間のオブジェクトを追跡する必要がある．そこで本研究では，BoT-SORT[18] を用いたトラッキング手法を導入する．BoT-SORT は，複数の判断基準を最適化することで，信頼性の低いオブジェクトも正確に追跡することが可能である．

本研究では，この YOLO-World と BoT-SORT を組み合わせることで，交通シーンに特化したマルチオブジェクトトラッキング（MOT）を実現する．その際，詳細に定義したクラスをプロンプトとして与える．

3.2 時空間シーングラフの構築

動画像が与えられたとき，前述した MOT を行い，検出されたオブジェクトの位置・クラス情報・追跡情報を用いて時空間シーングラフ G を式 (1) として構築する．

$$G = \{V, E\} \tag{1}$$

ここで，V はノード集合，E はエッジ集合である．

3.2.1 ノード集合 V の定義

MOT で得られた各オブジェクトをノード $v_i^t \in V$ として定義する．各フレーム毎に検出結果の参照を行い，オブジェクト毎の境界ボックス座標 B，クラスラベル c，ID 化されたトラッキング情報 id を取得する．MOT の出力を $\mathcal{D}_t$ と定義すると，式 (2) のように表される．

$$\mathcal{D}_t = \{(B_i^t, c_i^t, id_i^t) \mid i = 1, \ldots, N_t\} \tag{2}$$

ここで，N_t は各フレーム t 内の検出オブジェクト数，$B_i^t = (x_{min}, y_{min}, x_{max}, y_{max})$ はオブジェクト i の境界ボックス座標を表す．次に，境界ボックス座標の中心座標を求める．このとき，座標値の範囲が入力によって大きく変動するため，入力画像サイズの幅 w と高さ h を考慮した正規化中心座標 $\tilde{B}_i^t$ を式 (3) によって求める．

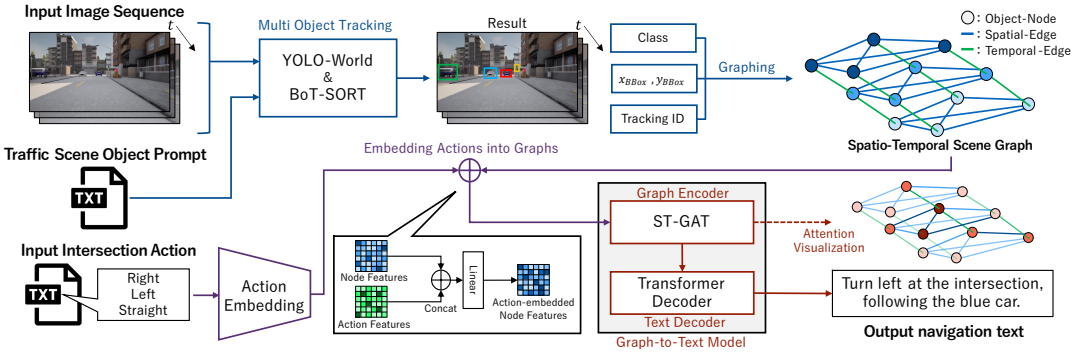


図 1: 提案手法のアーキテクチャ

$$\tilde{B}_i^t = \left(\frac{x_{min} + x_{max}}{2w}, \frac{y_{min} + y_{max}}{2h} \right) \tag{3}$$

これらの特徴を用いて，各ノード v_i^t の特徴ベクトル f_i^t を式 (4) のように付与する．

$$f_i^t = [\tilde{B}_i^t, e_{c_i^t}] \tag{4}$$

ここで，$e_{c_i^t}$ はクラスラベル c_i^t に対応する one-hot ベクトルである．

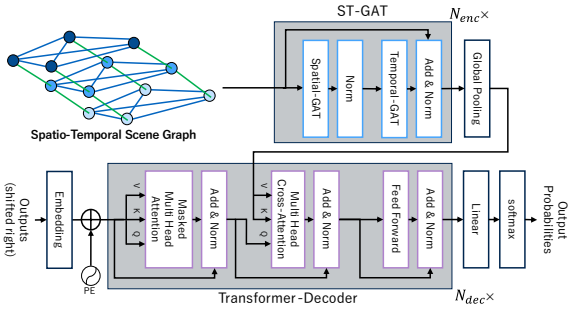


図 2: Graph-to-Text モデル

3.2.2 エッジ集合 E の定義

ノード同士を接続するエッジは，空間的エッジ E_{spatial} と時系列エッジ E_{temporal} で構成される．

空間的エッジ E_{spatial} は，同一フレーム内のオブジェクト間の関係性を示し，式 (5) のように定義する．

$$E_{\text{spatial}} = \{((v_i^t, v_j^t), w_{ij}^t)\}, w_{ij}^t = \| \tilde{B}_i^t - \tilde{B}_j^t \|_2 \tag{5}$$

ここで，w_{ij}^t はエッジに対する重みを示し，オブジェクト間のユークリッド距離を付与する．フレーム毎にオブジェクト間の距離は動的に変化するため，こうした特徴をグラフ全体で表現することを図る．

時系列エッジ E_{temporal} は，時系列方向におけるオブジェクト間の関係性を示し，式 (6) のように定義する．

$$E_{\text{temporal}} = \{((v_i^t, v_j^{t+1})) \mid id_i^t = id_j^{t+1}\} \tag{6}$$

この処理では，前後のフレームでトラッキング ID が一致するノードに対してエッジが接続される．

最終的に，エッジ集合 E は式 (7) となる．

$$E = E_{\text{spatial}} \cup E_{\text{temporal}} \tag{7}$$

3.3 Action のグラフへの埋め込み

本タスクでは，交差点における動作情報が明示的に与えられることを前提とする．したがって，交差点における動作情報 Action はテキスト形式で与えることを考える．事前にシーングラフに Action を埋め込むことで，Action に基づいて着目すべきノードが強調されるような効果を図る．ここで，Action の埋め込みベクトル $\tilde{a}$ を取得するため，式 (8) に示す操作を行う．

$$\tilde{a} = \text{Embedding}(a) \tag{8}$$

続いて，式 (9) に示すようにグラフのノード特徴 f_i^t と Action の埋め込みベクトル $\tilde{a}$ を統合する．

$$h_i^t = \sigma(\mathbf{W}_f[f_i^t, \tilde{a}] + \mathbf{b}_f) \tag{9}$$

ここで，h_t^i は Action を統合したノード特徴として新たに扱う．また，$\mathbf{W}_f$ は特徴統合のための線形変換行列，$\mathbf{b}$ はバイアス項，σ は活性化関数を示す．

3.4 Graph-to-Text モデル

本研究では，時空間シーングラフから文章の生成を行う Graph-to-Text モデルを提案する．モデルの詳細を図 2 に示す．Graph-to-Text モデルは，グラフの特徴抽出を行う Graph Encoder と文章生成を行う Text Decoder で構成される．

3.4.1 Graph Encoder

時空間シーングラフの特徴抽出を行うエンコーダモデルとして，空間方向と時系列方向に分けて Attention を適用する Spatial Temporal GAT (ST-GAT) を構築する．各ノード v_i^t の特徴量 h_t^i は，空間方向および時系

列方向の近傍ノード情報をそれぞれ GAT 層を通じて集約・更新することで得られる．いずれの層においても，Attention の計算は式 (10) の一般式に基づいて行う．

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}_{(*)i}} \alpha_{(*)ij}\mathbf{W}\mathbf{h}_j^{(l)}\right)$$
$$\alpha_{(*)ij} = \frac{\exp(\sigma(\theta^\top[\mathbf{W}\mathbf{h}_i^{(l)}, \mathbf{W}\mathbf{h}_j^{(l)}]))}{\sum_{k\in\mathcal{N}_{(*)i}} \exp(\sigma(\theta^\top[\mathbf{W}\mathbf{h}_i^{(l)}, \mathbf{W}\mathbf{h}_k^{(l)}]))} \quad (10)$$

ここで，$\mathcal{N}_{(i)}$ はノード v_i^t の近傍ノードの集合であり，$(*)$ は GAT 層の種類（spatial or temporal）を表す．また，$a_{(*)ij}$ は Attention の重み，θ は Attention 係数を計算するための学習可能な重みベクトルを示す．

ST-GAT による各ノード特徴量 h_t^i の更新後，グラフ全体の表現を求めるため，Global Average Pooling を適用し，情報集約を行う．計算式を式 (11) に示す．

$$z = \frac{1}{|V|}\sum_{i\in V} h_i^{(L)} \quad (11)$$

ここで，$|V|$ はノード数，$h_i^{(L)}$ は最終層 L におけるノード v_i^t の特徴量，z はグラフ全体の表現ベクトルである．

3.4.2 Text Decoder

Text Decoder では，Graph Encoder により抽出されたグラフ全体の特徴量 z から，Transformer Decoder を用いて文章の生成を行う．学習時，Transformer Decoder は，入力シーンに対応した案内文と抽出したグラフ特徴量を入力として，単語列を逐次的に予測し，データセットに含まれる文章の統計的な特徴を獲得する．学習時の動作を式 (12) に示す．

$$y_t = \text{TransformerDecoder}(z, y_{t-1}) \quad (12)$$

ここで，y_t は文字列のインデックス t における単語埋め込みベクトルである．推論時は，グラフ特徴量と文章の開始を示すトークンを入力とし，学習時に獲得した文章表現の特徴を基に，単語列を逐次的に生成する．

3.4.3 モデルの学習

Graph-to-Text モデルは End-to-End で学習を行う．学習の目的は，モデルが生成した文章と正解の文章との誤差 L を最小化することであり，式 (13) に示すように cross-entropy 損失を用いる．

$$L = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{T} targets_{i,j}\cdot\log(outputs_{i,j}) \quad (13)$$

ここで，M はバッチサイズ，T は最大シーケンス長，$targets_{i,j}$ は正解文の i 番目の文における j 番目の単語を表す one-hot ベクトル，$outputs_{i,j}$ はモデルが出力した対応する単語の予測確率である．

3.4.4 シーングラフの Attention 可視化

モデルの推論時，式 (10) より最終層 L における各エッジに対する Attention スコア $\alpha_{(spatial)ij}$，$\alpha_{(temporal)ij}$ をグラフ上に可視化することで，モデルの解釈を可能とする．ノード v_i の着目度（ノード Attention スコア）A_i は式 (14) より求められる．

$$A_i = \sum_{j\in\mathcal{N}(i)} \left(\alpha_{(spatial)ij} + \alpha_{(temporal)ij}\right) \quad (14)$$

式 (14) より A_i のスコアが高いほど案内文生成時に着目した重要なノードであると解釈できる．

4 データセット

本研究のナビゲーションタスクには，走行車両の車載カメラ映像と対応する案内文のペアからなるデータセットが必要となる．Waymo Open Dataset[19] など車載カメラの映像を用いたデータセットは多数存在するが，案内文のアノテーションは付与されていない．そこで，本研究では案内文生成に特化したデータセットをシミュレーション環境を用いて独自に作成する．

4.1 走行シーンの撮影

データセットの作成には，CARLA Simulator[20] を用いる．CARLA は車両・歩行者の生成や自動運転機能を備えており，本研究ではこれらの機能を活用し走行シーンを撮影した．撮影には 8 つのマップを用い，撮影条件は以下の通り設定する．

- フレームレート:：10 fps
- 天候条件：ClearNoon, WetNoon
- 撮影範囲：交差点約 50m 手前から交差点通過直後
- 選定基準：注目オブジェクトが 1 つ以上存在

4.2 案内文のアノテーション

案内文のアノテーションは手動で実施する．表現の統一性を確保するため，案内文の先頭に動作情報（交差点での進行方向）を記述し，注目対象に基づいた案内文を作成する．図 3 に，作成したデータの例を示す．

4.3 作成したデータセットの概要

作成したデータセットは，合計 160 シーン，計 10,219 フレームで構成される．各シーンには，前述した案内文，進行方向における動作情報（テキスト形式）が含まれる．また，データセットの構成として，学習用を 136 シーン，評価用を 24 シーンとする．

図 3: 撮影シーンと作成した案内文の例

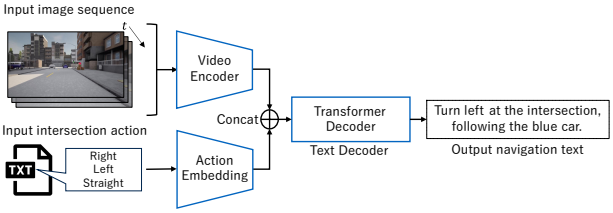


図 4: ベースライン手法の概要

5 評価実験

本章では，評価実験を通じて提案手法の有効性を検証する．本実験では，動画像から直接特徴量を抽出する手法をベースライン手法とし，提案手法との比較を行う．また，モデルへの入力フレーム数を 5 フレーム，10 フレーム，15 フレームに設定し，異なるフレーム長が案内文の生成精度に与える影響について分析する．各モデルで生成された案内文の精度を評価する際，BLEU[21]，METEOR[22]，ROUGE[23] を評価指標として用いる．

5.1 ベースライン手法

ベースライン手法として動画像から直接特徴量を抽出する手法を用いる．ベースライン手法として採用するアーキテクチャの概要を図 4 に示す．本実験では，Video Encoder として，CNN ベースの 3DCNN[10] と 3DResNet[24]，動画像の各フレーム画像を CNN で処理し時系列処理を Transformer で行う Video Transformer Network (VTN)[25]，動画像の空間・時系列方向を全て Transformer で処理する Video Vision Transformer (ViViT)[26] を用いる．VTN と ViViT においては，動画分類として用いられるモデルのため，最終層を取り除いたモデルを Video Encoder として用いる．また，走行シーンの動画像とペアとして入力される交差点における動作情報 Action については，埋め込みベクトル化後，Video Encoder の出力と結合する形式とする．

5.2 実験条件

提案手法及びベースライン手法に適用するハイパーパラメータを表 1(a)，表 1(b) に示す．両手法の比較での公平性を保つため，Text Decoder の設定は同一の値とする．それに従い，Graph Encoder および Video Encoder の出力次元数も同一とする．

学習における設定は，学習率を 1.0×1.0^{-4}，エポック数を 100，Dropout 率を 0.3，バッチサイズを 32 とする．学習の最適化アルゴリズムには AdamW を用いる．これらの設定は，提案手法およびベースライン手法の全てのモデルで統一する．

表 1: ハイパーパラメータ設定

(a) 提案手法

Module	Model	Hyperparam.	Value
Action Embedding	Embedding	Output dim	128
Feature Fusion	Linear	Output dim	256
Graph Encoder	ST-GAT	Hidden dim Output dim heads layers	256 1024 4 3
Text Decoder	Transformer Decoder	FFN dim heads layers	2048 8 6

(b) ベースライン手法

Module	Model	Hyperparam.	Value
Action Embedding	Embedding	Output dim	1024
Video Encoder	① 3DCNN ② 3DResNet ③ VTN ④ ViViT	Output dim	1024
Feature Fusion	Linear	Output dim	1024
Text Decoder	Transformer Decoder	FFN dim heads layers	2048 8 6

5.3 定量的評価

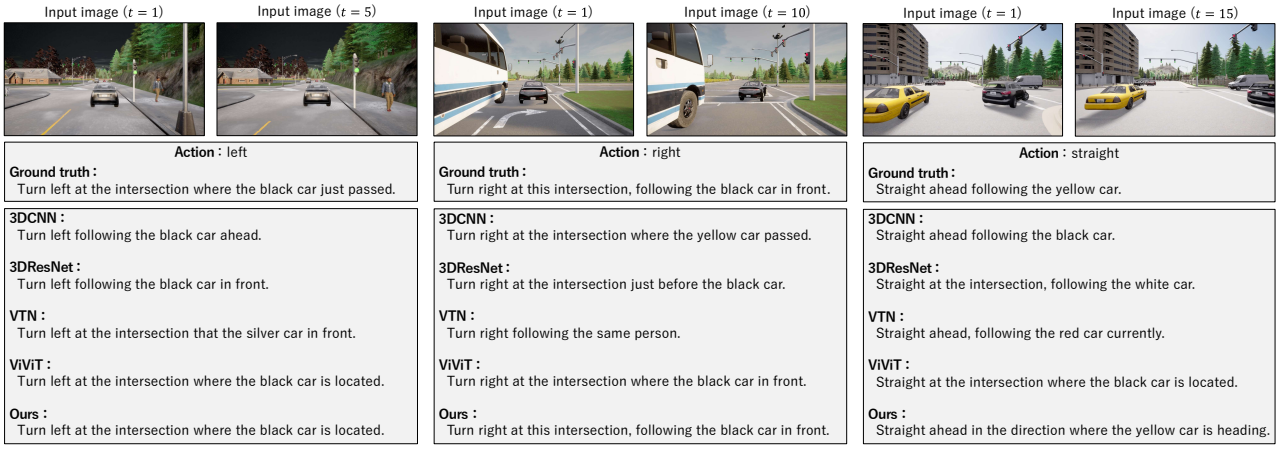
提案手法およびベースライン手法の各モデルで生成された案内文の精度について定量的評価によって比較を行う．評価結果を表 2 に示す．表 2 における表記について，Method はベースラインにおいては Video Encoder で用いたモデル名とする．また，スコア表記においては，太字は各フレーム数において最も高精度なもの，アンダーラインは全体で最も高精度なものを示す．

まず，従来手法である 3DCNN，3DResNet，VTN においては，フレーム数が増加するにつれて精度が低下する傾向がある．特に，3DCNN と VTN は 15 フレーム時にスコアが顕著に低下しており，時間的な情報を長期間統合することが困難であることが示唆される．一方で，ViViT はフレーム数が増加しても精度が向上する傾向にある．ViViT は，Transformer ベースのアーキテクチャを採用しており，フレーム間の長期的な依存関係を Self-Attention 機構によって適切に処理できる．これにより，フレーム数が増加した場合においても，時系列情報を捉えやすく，安定した案内文生成が可能になっていると考えられる．また，提案手法（Ours）は全てのフレーム数において他の手法を上回る精度を達成しており，フレーム数が増加するほどより顕著に精度が向上していることが確認できる．特に，15 フレーム時には全てのスコアで最高の精度を達成している．これは，時空間シーングラフの導入により，フレーム間のオブジェクト関係を適切に捉え，重要な情報を強調することができたためと考えられる．

表 2: 各モデルで生成された案内文の精度結果

Method	5 frame				10 frame				15 frame			
	B-1	B-4	M	R	B-1	B-4	M	R	B-1	B-4	M	R
3DCNN	0.568	0.322	0.575	0.643	0.551	0.292	0.538	0.617	0.519	0.268	0.515	0.601
3DResNet	0.459	0.197	0.446	0.559	0.448	0.173	0.439	0.547	0.457	0.180	0.449	0.534
VTN	0.583	0.337	0.578	0.565	0.412	0.142	0.378	0.537	0.379	0.099	0.377	0.471
ViViT	0.524	0.266	0.538	0.592	0.540	0.285	0.551	0.603	0.549	0.274	0.559	0.611
Ours	**0.610**	**0.363**	**0.635**	**0.668**	**0.617**	**0.382**	**0.646**	**0.675**	**0.631**	**0.388**	**0.649**	**0.677**

※ B-1 : BLEU-1, B-4 : BLEU-4, M : METEOR, R : ROUGE



(a) 入力フレーム数：5　(b) 入力フレーム数：10　(c) 入力フレーム数：15

図 5: 各モデルの案内文生成結果

5.4 定性的評価

提案手法およびベースライン手法の各モデルで生成された案内文について定性的に評価を行い，両手法を比較することで有効性の検証を行う．

各モデルの案内文生成結果を図 5 に示す．入力フレーム数が 5 の場合（図 5(a)），VTN を除く全て手法において “black car” を注目とした案内文を生成しており，適切な説明を行っていることが確認できる．VTN においては注目対象が “silver car” となっているが，車両の色特徴が “black car” と近いためだと推察される．入力フレーム数が 10 の場合（図 5(b)），3DCNN および VTN においては入力画像内に存在しないオブジェクトを中心とした案内文を生成しており，オブジェクトを捉えられていないことが確認できる．一方で，3DResNet，ViViT，提案手法においては Groud Truth と同様の “black car” を中心とした案内文を生成できているが，提案手法が最も適切な説明をしていることが確認できる．入力フレーム数が 15 の場合（図 5(c)），Groud Truth と同様の “yellow car” を中心とした案内文を生成できているものは提案手法のみであり，最も適切な説明となっている．ベースライン手法においては，最も動作の変化が大きい “black car” もしくは画像内に存在しないオブジェクトを注目しており，不適切な説明となっている．長期のフレームにおいては提案手法が有効的であることを示唆している．

次に，提案手法における案内文生成において，推論時の Graph Encoder から取得した Attention を時空間シーングラフ上に可視化する．Attention の可視化結果を図 6 に示す．上段の例では，最も着目しているノードは 8 フレーム目の “white car” であり，生成案内文の着目しているオブジェクトと一致する．下段の例においても，1 フレーム目の “black car” に着目しており，生成案内文の着目しているオブジェクトと一致する．また，どのオブジェクト同士の関係性に着目しているかについても，Edge Attention により確認できる．これらの結果より，提案手法はモデルが生成した案内文の判断根拠をシーングラフを通して視覚的に説明可能であることを確認できる．

6 おわりに

本研究では，車両の視界情報を基に動的な環境を理解し，運転手に直感的で自然な案内文を生成する手法を提案した．具体的には，走行シーンのオブジェクト関係を時空間シーングラフとして構築し，Graph-to-Text モデルを用いて案内文を生成した．特に，GAT による重要情報の強調と，Attention の可視化を通じたモデルの判断根拠の明確化を図った．評価実験では，提案手法が既存の CNN や Transformer ベースの Video Encoder

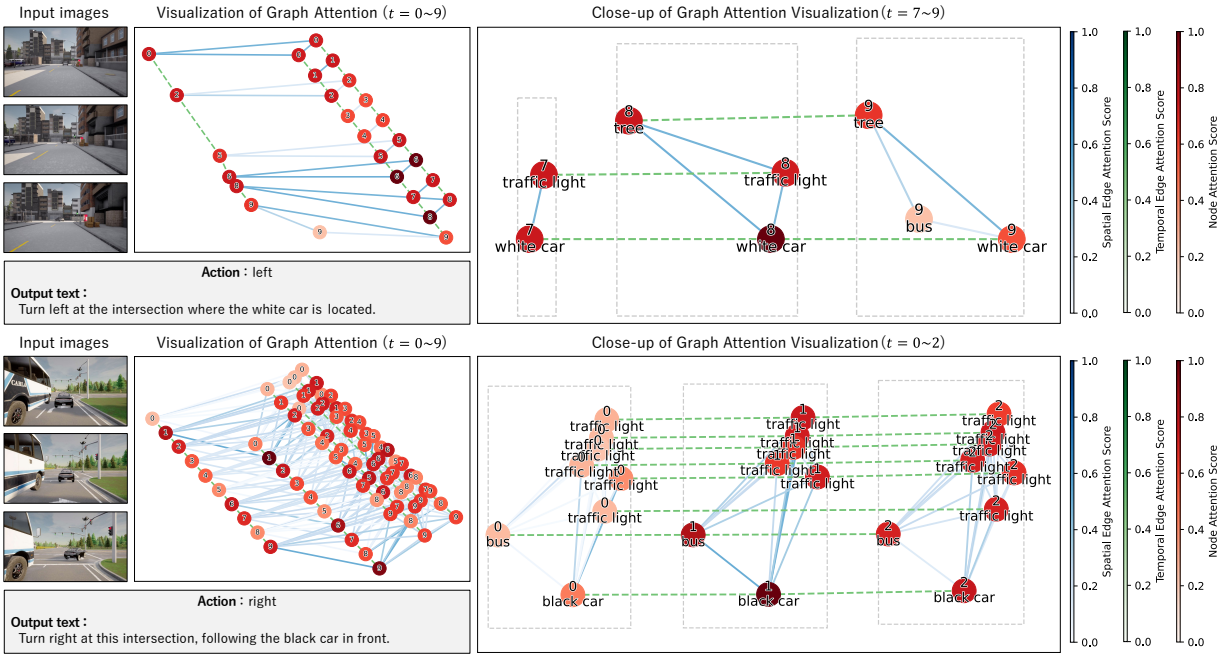


図 6: Graph Encoder で得られた Attention の可視化結果．グラフは全体の中から 10 フレーム分を抜粋し，さらに着目度が高いと判断される 3 フレームを強調した結果を示す．ノードに表示されるラベルは，対応するフレーム番号および MOT により得られたオブジェクトのクラス名を示す．

を用いたベースライン手法よりも一貫して高い精度を達成し，フレーム数の増加に伴う精度向上が確認された．特に，長期間の情報統合において，時空間シーングラフが有効に機能し，適切な案内文の生成を可能にした．定性的評価においても，提案手法はより適切な対象を注目し，正確な説明を行っていることを確認した．また，モデルから得られた Attention を可視化した結果より，モデルが生成した案内文とグラフ上で着目しているオブジェクトが一致しており，視覚的説明を実現した．これにより，モデルの信頼性を高めることが可能であることを示した．

今後の課題として，より複雑な環境における適応や，多様な運転シナリオへの適用性の向上が挙げられる．したがって，今後はデータセットの増強による提案手法の有効性について検証を行う予定である．

参考文献

[1] Anil K. Kandangath and Xiaoyuan Tu. Humanized navigation instructions for mapping applications, April 2015. US Patent application.

[2] Amanda Cercas Curry, Dimitra Gkatzia, and Verena Rieser. Generating and evaluating landmark-based navigation instructions in virtual environments. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 90–94, 2015.

[3] Hayato Suzuki, Kota Shimomura, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi, Shota Okubo, Nanri Takuya, and Wang Siyuan. Human-like guidance by generating navigation using spatial-temporal scene graph. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 1988–1995, 2024.

[4] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[5] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[7] Xinxin Zhu, Lixiang Li, Jing Liu, Haipeng Peng, and Xinxin Niu. Captioning transformer with stacked attention modules. *Applied Sciences*,

8(5):739, 2018.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[9] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. Cptr: Full transformer network for image captioning, 2021.

[10] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.

[11] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *CVPR*, 2022.

[12] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[13] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 5425–5434, 2017.

[14] Shijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[15] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 690–706, 2018.

[17] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024.

[18] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking, 2022.

[19] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, and Zhifeng Chen. The waymo open dataset: High resolution sensor data for autonomous driving. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] Alexey Dosovitskiy, Thomas Brox, Nikolay Stoyanov, Tom Erez, Eddy Ilg, Alexander Pishchulin, Maxim Yankovskiy, and Daniel Cremers. Carla: An open-source simulator for autonomous driving, 2017.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL)*, 2002.

[22] Michael Denkowski and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014.

[23] Chin-Yew Lin. Rouge: A package for automatic evaluation of summarization quality. *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004.

[24] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[25] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3156–3165, 2021.

[26] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.