GENNAV: 移動指示文に基づく目標領域へのポリゴンマスク生成

勝又 圭[1,a)]　飯岡 雄偉[1,b)]　細見 直希[2,c)]　翠 輝久[3,d)]　山田 健太郎[2,e)]　杉浦 孔明[1,f)]

概要

本研究では, モビリティから取得した前方カメラ画像と自然言語指示文から対象領域の位置を特定するタスクを扱う. 本タスクは, 境界が曖昧な stuff 型の対象領域に対して, 存在有無の予測とセグメンテーションマスク生成の双方を行うため困難である. そこで本研究では, 対象領域の存在有無を明示的に予測しポリゴンベースのセグメンテーションマスクを生成するモデル GENNAV を提案する. また, GENNAV を評価するため, no-target/single-target/multi-target の３種を含む新たなベンチマーク GRiN-Drive ベンチマークを構築した. 実験の結果, GENNAV は標準的な評価指標においてベースライン手法を上回った. さらに, 5 つの都市において実環境実験を実施し, ゼロショット性能を検証した結果, GENNAV の実環境における実用性が確認された.

1. はじめに

公共交通機関が脆弱化しつつある地域では, 住民の移動手段を確保することが重要な社会的課題となっている. このような背景から, 新たな移動手段としてパーソナルモビリティの開発が数多く試みられている [5, 3]. しかし既存技術の多くは, 多様なシーンでの柔軟な対応には不十分であり, 広範なユーザ層が直感的かつ簡便に操作可能なインターフェースが求められているため, 移動指示文の適切な理解は極めて重要である.

そこで本研究では, モビリティから取得した前方カメラ画像とユーザーの与えた自然言語移動指示文から, ユーザーが意図する対象領域を特定するタスクを扱う. 実際の運転シーンではモビリティの移動によりランドマークが画像外に移動するケースのように, 指示に対応する領域が必ずしも存在するとは限らない. このように画像内に指示に対応する領域が存在しない場合, モデルは “no-target” を予測することが求められる. 図 1 は本タスクの代表例を示す. ユーザーが “Please park to the left of the black car on the right” と指示した場合, 黒い車の左側領域に対応するマスクを生成する必要がある. 一方, 利用者が画像内に存在しない赤い車をランドマークとして “Park left of the red car on the right” と指示した場合, モデルは “no-target” を予測することが期待される.

本タスクは, 対象領域の存在有無を予測すると同時に, その領域へセグメンテーションマスクを生成する必要があるため困難である. 特に, 明確な境界を持つ thing 型の領域 (例：歩行者, 標識, 車両) と異なり stuff 型の領域 (例：道路) には明確な境界がないため本タスクは難しいと言える. そこで本研究では, 対象領域の存在有無を明示的に予測し, stuff 型の対象領域に対してセグメンテーションマスクを生成する手法 GENNAV を提案する.

本研究の主な貢献は以下である.

- 対象領域の有無を予測し, ポリゴンベースのセグメンテーションマスクを生成する Existence Aware Polygon Segmentation Module を導入する.
- ランドマークの空間分布に基づいて画像をパッチ化し, 潜在的ランドマークに関連する視覚表現を取得する Landmark Distribution Patchification Module を導入する.
- No-target/single-target/multi-target の 3 種類を含むベンチマーク GRiN-Drive を構築する.
- 5 つの都市における実環境実験により, GENNAV のゼロショット性能を検証する.

2. 問題設定

本研究では, モビリティから取得した前方カメラ画像とユーザーの自然言語指示に基づき, 対象領域の存在有無と位置を同時に予測する Generalized Referring Navigable Regions (GRNR) タスクを扱う. 本タスクでは, 移動指示文で指定された対象領域が入力画像内に存在するか否かを判定することが求められる. また, 対象領域が一つ以上

1 慶應義塾大学
2 株式会社本田技術研究所 先進技術研究所
3 Honda Research Institute USA
a) ke59ka77@keio.jp
b) kmngrd1805@keio.jp
c) naoki_hosomi@jp.honda
d) tmisu@honda-ri.com
e) kentaro_yamada@jp.honda
f) komei.sugiura@keio.jp

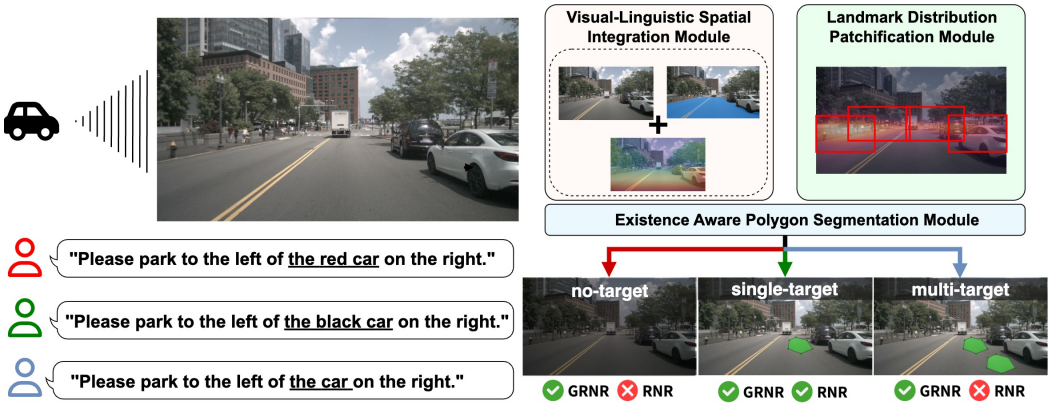

図 1: GRNR タスクの代表例.

存在する場合，各対象領域に対してセグメンテーションマスクを生成する必要がある．図 1 に GRNR タスクの代表例を示す．Referring Navigable Regions タスクと異なり，GRNR タスクでは複数の対象領域や対象がない場合を含む任意の数の対象領域を指定する指示文が含まれる．本タスクでは，実運用において既存の下位モジュールが十分に軌道予測を行えるため，対象領域への軌道予測は扱わない．

本論文で用いる用語を以下に定義する．“移動指示文” とは，モビリティを目的地へ誘導するための自然言語の指示文を指す．“対象領域” とは，与えられた移動指示文によって指定される領域を指す．また，“ランドマーク” とは，対象領域を指定する際に参照される物体のことである．

3. 提案手法

本研究ではポリゴンベースのセグメンテーションを対象領域の有無判定を行うモデル GENNAV を提案する．図 2 に GENNAV のモデル図を示す．GENNAV は，Landmark Distribution Patchification Module（LDPM），Visual-Linguistic Spatial Integration Module（VLSiM）および Existence Aware Polygon Segmentation Module（ExPo）から構成される．モデルへの入力 $\boldsymbol{x}$ は $\boldsymbol{x} = \{\boldsymbol{x}_{\text{img}}, \boldsymbol{x}_{\text{inst}}\}$ であり，$\boldsymbol{x}_{\text{img}} \in \mathbb{R}^{3\times W\times H}$，$\boldsymbol{x}_{\text{inst}} \in \{0,1\}^{V\times L}$ はそれぞれ前方カメラ画像および移動指示文を表し，W, H, V, および L は画像幅, 画像の高さ, 語彙サイズ，および最大トークン長を示す．$\boldsymbol{x}_{\text{inst}}$ は GISTEmbed [6] により前処理を行い，言語特徴量 $\boldsymbol{h}_{\text{inst}} \in \mathbb{R}^{C_{\text{inst}}}$ を得る．ここで C_{inst} は言語特徴量の次元数を表す．

3.1 LDPM

本モジュールは，ランドマークの空間分布に基づいて効率的にパッチを分割し，潜在的ランドマークに関連する視覚表現を取得する．本タスクにおいて $\boldsymbol{x}_{\text{inst}}$ は歩行者や車両などのランドマークの理解が重要である．しかし，多くの既存手法（例： [5, 3]）は計算コスト削減を目的に入力画像を低解像度にリサイズするため，遠方のランドマークに割り当てられる画素数が不足し，ランドマークの詳細を捉えることが困難である．よって，本モジュールでは効率的に高解像度の画像を活用するため，パッチ化を行う．ただし，均一な分割では空など余分な領域を多く含むため，ランドマークが密集しやすい領域を重点的にカバーするランドマーク分布ベースのパッチ化を行う．図 2 の LDPM（緑背景）の左に示す図に，実際のパッチ分割の領域（赤枠）と Talk2Car データセット [2] の訓練集合におけるランドマークの分布を示す．

本モジュールの入力は $\boldsymbol{x}_{\text{img}}$ であり，前述のようにパッチ化したのち，それぞれを DINOv2 [4] で特徴量を取得し，CNN に入力する．そして，得られた特徴量を加算して最終的な特徴量 $\boldsymbol{h}_{\text{ldp}}$ を得る．

3.2 VLSiM

本モジュールは，$\boldsymbol{x}_{\text{img}}$，$\boldsymbol{h}_{\text{inst}}$，擬似深度画像，および路面領域を統合することで，言語情報と空間情報の関係をモデル化する．VLSiM への入力は $\boldsymbol{x}_{\text{img}}$ および $\boldsymbol{h}_{\text{inst}}$ である．はじめに，DINOv2 と CNN を用いて $\boldsymbol{x}_{\text{img}}$ から視覚特徴量を抽出する．この手順を $f_{\text{vis}}(\boldsymbol{x}_{\text{img}})$ と表す．

次に，単眼深度推定モデル（例：Depth Anything V2 [4]）を用いて $\boldsymbol{x}_{\text{img}}$ から擬似深度画像を生成し，これを $\boldsymbol{x}_{\text{img}}$ と重畳する．その画像から DINOv2 と CNN を用いて特徴量抽出を行う．これを $f_{\text{depth}}(\boldsymbol{x}_{\text{img}})$ とする．

最後に，$\boldsymbol{h}_{\text{inst}}$，$f_{\text{vis}}(\boldsymbol{x}_{\text{img}})$，および $f_{\text{depth}}(\boldsymbol{x}_{\text{img}})$ を統合する．さらに，車両の侵入不可領域（例：空や遮蔽物の存在する領域）へのマスク生成を抑制するため，$f_{\text{depth}}(\boldsymbol{x}_{\text{img}})$ と路面領域を示すマスクを統合する．路面マスクは $\boldsymbol{x}_{\text{img}}$ を入力としてセマンティックセグメンテーションモデル（例：PIDNet [8]）により取得し，これを CNN に入力する．これを $f_{\text{road}}(\boldsymbol{x}_{\text{img}})$ とする．最終的な VLSiM の出力として，$\boldsymbol{h}_{\text{mm}} = \boldsymbol{h}_{\text{inst}} \odot \{f_{\text{vis}}(\boldsymbol{x}_{\text{img}}) + f_{\text{depth}}(\boldsymbol{x}_{\text{img}})\} \odot \{f_{\text{road}}(\boldsymbol{x}_{\text{img}}) + f_{\text{depth}}(\boldsymbol{x}_{\text{img}})\}$ を得る．

3.3 ExPo

本モジュールは $\boldsymbol{h}_{\text{inst}}$，$\boldsymbol{h}_{\text{mm}}$，$f_{\text{road}}(\boldsymbol{x}_{\text{img}})$，および $\boldsymbol{h}_{\text{ldp}}$ を統合し，対象領域の有無を予測するとともにポリゴンベースのセグメンテーションマスクを生成する．本モジュールは分類ヘッド (存在判定) と回帰ヘッド (マスク生成) の二つのヘッドから構成される．特に回帰ヘッドはポリゴ

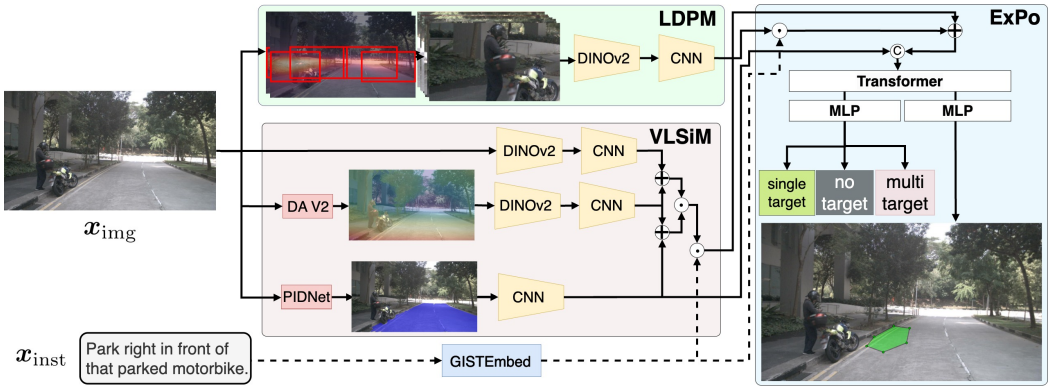

図 2: GENNAV のモデル図．DA は Depth Anything [9] を表す．

ンベースの手法を用いることで既存のピクセルベース手法 [5, 3] より効率的な推論を実現する（詳細は 4.4 節を参照）．本モジュールは $\boldsymbol{h}_{\text{inst}}$，$\boldsymbol{h}_{\text{mm}}$，$f_{\text{road}}(\boldsymbol{x}_{\text{img}})$ および $\boldsymbol{h}_{\text{ldp}}$ を入力とし，$\boldsymbol{h}_{\text{cmm}} = (\boldsymbol{h}_{\text{ldp}} \odot \boldsymbol{h}_{\text{inst}}) + \boldsymbol{h}_{\text{mm}}$ を得る．回帰ヘッドは $\hat{\boldsymbol{c}}_i = \text{MLP}([\boldsymbol{h}_{\text{cmm}}; f_{\text{road}}(\boldsymbol{x}_{\text{img}})])$ を用いてポリゴン頂点 $\hat{\boldsymbol{c}}_i\,(i = 1, \ldots, n_{\text{pt}})$ を予測する．分類ヘッドにおいても同様に no-target/single-target/multi-target に対する予測確率 $p(\hat{\boldsymbol{y}}_{\text{ex}})$ を出力する．よって ExPo の出力は $\hat{\boldsymbol{y}} = \{p(\hat{\boldsymbol{y}}_{\text{ex}}), \hat{\boldsymbol{c}}_i\}$ となる．本手法における損失関数には L1 損失およびクロスエントロピー損失を用いる．

4. 実験

4.1 実験設定

モデルの性能を GRNR タスクにおける 3 種類のシナリオで評価するため Talk2Car-RegSeg [5] および Refer-KITTI-V2 [11] を基に新しく GRiN-Drive ベンチマークを構築した．Talk2Car-RegSeg では移動指示文と対象領域へのマスクが付与された前方カメラ画像のペアを含むが，no-target および multi-target のサンプルが存在しない．よって我々は no-target および multi-target サンプルを新規収集し，GRiN-Drive を構築した．GRiN-Drive は 17,114 サンプルの画像，移動指示文，および対象領域ごとの 0 個以上のセグメンテーションマスクから構成される．語彙サイズ，語数総計，および平均文長はそれぞれ 2,699, 168,640, および 10.55 語である．訓練集合，検証集合，およびテスト集合はそれぞれ 14,973，1,413，および 758 サンプルである．

ベースライン手法として，ピクセルベース手法 (LAVT[10]，TNRSM [3]) と MLLM 手法 (GPT-4o [1], Qwen2-VL [7]) を採用した．評価尺度には msIoU，P@K，精度 (Acc.), および推論時間を用い，主要評価尺度は msIoU とした．我々は single-target/multi-target/no-target を公平に評価するため，新たな指標 msIoU を提案する．従来指標では，正しい no-target の予測に IoU=1.0 を与える一方で正確なマスク生成には 1.0 未満しか与えず，バイアスが生じてしまう．その結果，対象領域の有無の分類に偏った評価が行われる．実際，常に “no target” と予測する場合において既存評価手法の gIoU では 0.33 となる．このスコアは GRiN-Drive テスト集合における人間の性能である msIoU=0.17 を上回る．よって，msIoU は IoU が閾値 K を超えるサンプルに 1 を割り当て，閾値未満は K に基づき正規化する．さらに sIoU@k（$k = 1, \ldots, 1/K$）を平均することで，msIoU は対象サンプルと no-target サンプルを公平に評価する．msIoU は以下のように定義される．

$$\text{msIoU} = \text{mean}\left(\frac{1}{N}\sum_{i=1}^{N}\text{sIoU@}k_i\right),$$

$$\text{sIoU@}k_i = \begin{cases} \min\left(k \cdot \text{IoU}(\hat{y}_i, y_i), 1\right) & \text{if TP,} \\ 1 & \text{if TN,} \\ 0 & \text{if FP or FN.} \end{cases}$$

ここで，$\hat{y}_i$ と y_i はそれぞれ i 番目のサンプルに対する予測マスクと正解マスクを表す．また，N はサンプル数を示し，IoU$(\cdot)$ は両マスク間の IoU を表す．

表 1: ベースライン手法と提案手法の定量的比較結果

手法	解像度	Type	msIoU [%]↑	P@0.1 [%]↑	Acc. [%]↑	推論時間 [ms/sample]↓
LAVT [10]	224 × 224	Pixel	31.73 ±2.74	30.28 ±3.16	40.95 ±1.56	**9.29** ±0.24
TNRSM [3]	224 × 224	Pixel	37.90 ±0.82	44.05 ±1.33	67.23 ±2.72	502.69 ±0.22
LAVT [10]	640 × 640	Pixel	34.09 ±1.53	7.90 ±17.66	38.66 ±10.69	11.51 ±0.33
TNRSM [3]	640 × 640	Pixel	22.84 ±8.77	38.57 ±4.25	67.89 ±1.11	503.68 ±0.20
GPT-4o [1]	1600 × 900	Bbox	23.41 ±5.72	5.04 ±1.37	62.44 ±9.35	3525.68 ±701.04
Qwen2-VL [7]	1600 × 900	Bbox	24.06 ±1.15	3.85 ±0.95	64.25 ±0.91	1768.99 ±0.41
Qwen2-VL [7]	1600 × 900	Polygon	12.16 ±4.86	0.08 ±0.18	44.66 ±5.97	1771.41 ±6.20
GENNAV (ours)	640 × 640	Polygon	**46.35** ±1.52	**49.60** ±1.12	**75.41** ±0.67	31.31 ±0.05
被験者実験	1600 × 900	Polygon	56.08	56.00	88.00	-

4.2 定量的比較結果

表 1 に，GRiN-Drive ベンチマークにおけるベースライン手法，GENNAV，および人間の性能の定量的結果を示す．表中の値は 5 回の施行における平均と標準偏差である．“Type” は，各手法のセグメンテーション方式を示す．表 1 から，GENNAV の msIoU は 46.35%であり，ベースライン手法の中で最良の結果を得た TNRSM（224 × 224）を 8.45 ポイント上回った．同様に，GENNAV は P@0.1 で 49.60%，精度で 75.41%であり，ベースライン手法における最良のスコアと比較して，それぞれ 5.55 ポイントおよび 7.52 ポイント上回った．GENNAV の 1 サンプルあ

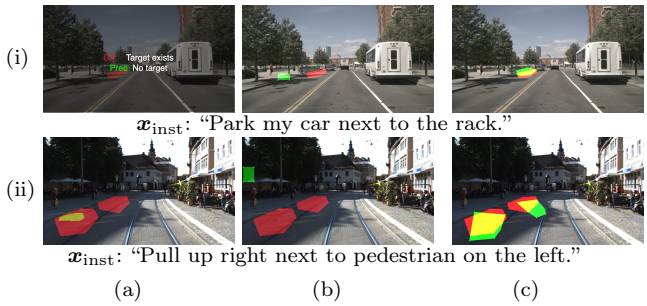

図 3: 定性的結果. 列 (a), (b) 及び (c) は TNRSM, Qwen2-VL (bbox) 及び GENNAV の結果を示す.

表 2: Ablation study の定量的結果.

モデル	Type	LDPM	TIaLFM		msIoU [%]↑	P@0.1 [%]↑	Acc. [%]↑
			f_{sp}	f_{road}			
(i)	polygon		✓	✓	43.05 $\pm$ 0.72	40.75 $\pm$ 1.77	74.62 $\pm$ 0.86
(ii)	polygon	✓		✓	44.13 $\pm$ 0.26	48.14 $\pm$ 0.44	74.01 $\pm$ 0.79
(iii)	polygon	✓	✓		44.53 $\pm$ 0.78	46.59 $\pm$ 2.56	73.75 $\pm$ 0.49
(iv)	pixel	✓	✓	✓	35.26 $\pm$ 1.21	27.18 $\pm$ 2.10	62.93 $\pm$ 2.03
(v)	polygon	✓	✓	✓	**46.35** $\pm$ 1.52	**49.60** $\pm$ 1.12	**75.41** $\pm$ 0.67

たりの推論時間は 31.31 ms であり,LAVT を除くすべてのベースライン手法より高速であった．また，LAVT には劣るものの，他の指標では大幅に上回った．msIoU，P@0.1，および精度における GENNAV との性能差は統計的に有意であった（$p < 0.05$）.

4.3 定性的結果

図 3 は GRiN-Drive ベンチマークにおける GENNAV とベースライン手法の定性的結果を示す．図 3 (i) は single-target サンプル，図 3 (ii) は multi-target サンプルの例である．図 3 (i) では，$\boldsymbol{x}_{\text{inst}}$ として “Park my car next to the rack” が与えられた．画像の中央左に自転車ラックがあるため，ラックの手前領域にマスクを生成する必要がある．GENNAV は適切にマスクを生成した一方で，TNRSM は “no-target” と予測し，Qwen2-VL(bbox) は不適切なマスクを生成した．図 3 (ii) の例では，$\boldsymbol{x}_{\text{inst}}$ は “Pull up right next to pedestrian on the left.” であった．モデルは道路左側に位置する各歩行者の横の領域にマスクを生成する必要がある．GENNAV は 2 人の歩行者それぞれに対して適切なマスクを生成した．しかし，ベースライン手法はいずれも適切なマスクを生成できず，特に TNRSM は誤って単一領域のみを予測した.

4.4 Ablation Study

各モジュールの有効性を検証するために ablation study を行った．表 2 にその結果を示す.

LDPM Ablation. LDPM の有効性を検証するために LDPM を除去し実験を行った．モデル (i) は msIoU においてモデル (v) と比べて 3.30 ポイント低下した．これは，LDPM がランドマーク分布に基づくパッチ分割によってランドマーク理解向上へ寄与していること示唆している.

VLSiM Ablation. $f_{\text{depth}}(\cdot)$ と $f_{\text{road}}(\cdot)$ の寄与を検証した．$f_{\text{depth}}(\cdot)$ を除去したモデル (ii) では msIoU が 2.22 ポイント低下した．これは，$f_{\text{depth}}(\cdot)$ が言語情報と空間情報の整合に寄与していることを示唆している．同様に $f_{\text{road}}(\cdot)$ を除去したモデル (iii) では，それぞれ 1.82 ポイント低下し，$f_{\text{road}}(\cdot)$ が移動不可領域の理解を高め，マスク生成性能を改善していることが示唆された.

(a) $\boldsymbol{x}_{\text{inst}}$: “Please go near the blue car traveling in the same direction.”

(b) $\boldsymbol{x}_{\text{inst}}$: “Stop to the right of the pedestrian on the left side.”

図 4: 実世界実験における GENNAV の定性的結果.

Decoder Ablation. Decoder 構造の影響を分析するため，ポリゴンベースとピクセルベースの手法を比較した．ピクセルベースの Decoder を採用したモデル (iv) は，msIoU においてモデル (v) を 11.09 ポイント下回った．さらに，モデル (iv) の推論時間は 47.92 ms/サンプルであり，モデル (v) の 31.31 ms/サンプルより遅かった．本結果から，ポリゴンベースの手法は計算コストを削減し，リアルタイム応用により適した構造であることが確認された.

4.5 実環境実験

ゼロショットの実環境実験において GENNAV の性能を検証した．実環境実験では，5 都市にて，計 20 本の動画を収集した．動画データは，トンネルや積雪路など，多様な交通・環境条件下で収集し，実環境における様々な条件のサンプルを収集した．図 4 は，実環境実験における定性的結果を示す．図 4 (i) は single-target，図 4 (ii) は multi-target の例である．図 4 (i) では，$\boldsymbol{x}_{\text{inst}}$ は “Please go near the blue car traveling in the same direction.” であり，GENNAV は同じ方向に走行する青い車の後方領域にマスクを適切に生成した．図 4 (ii) の例では，$\boldsymbol{x}_{\text{inst}}$ が “Stop to the right of the pedestrian on the left side.” であった．モデルは道路左側にいる各歩行者の右側領域にマスクを生成する必要がある．この例でも GENNAV は 2 人の歩行者それぞれに対して適切にマスクを生成した.

5. おわりに

本論文では GRNR タスクを扱い，stuff 型の対象領域に対して有無を判定し，セグメンテーションマスクを生成する GENNAV を提案した．本研究の貢献は次のとおりである．3 つの新規モジュール ExPo，LDPM，および VLSiM から構成される GENNAV を導入した．次に，single-target/no-target/multi-target の 3 種類のサンプルを含む GRiN-Drive ベンチマークを構築した．提案手法 GENNAV は GRiN-Drive ベンチマークにおいてベースライン手法を上回った．また，実環境実験を通じてゼロショット設定下で実環境における実用性を確認した.

参考文献

[1] Achiam, J., Adler, S., Agarwal, S. et al.: GPT-4 Technical Report, *arXiv preprint arXiv:2303.08774* (2023).

[2] Deruyttere, T., Vandenhende, S., Grujicic, D. et al.: Talk2Car: Taking Control of Your Self-Driving Car, *EMNLP*, pp. 2088–2098 (2019).

[3] Hosomi, N., Hatanaka, S. et al.: Trimodal Navigable Region Segmentation Model: Grounding Navigation Instructions in Urban Areas, *IEEE RA-L*, Vol. 9, No. 5, pp. 4162–4169 (2024).

[4] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W. et al.: DINOv2: Learning Robust Visual Features without Supervision (2023).

[5] Rufus, N., Jain, K., Nair, K., Gandhi, V. and Krishna, M.: Grounding Linguistic Commands to Navigable Regions, *IROS*, pp. 8593–8600 (2021).

[6] Solatorio, A.: GISTEmbed: Guided In-sample Selection of Training Negatives for Text Embedding Fine-tuning, *arXiv preprint arXiv:2402.16829* (2024).

[7] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R. et al.: Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution, *arXiv preprint arXiv:2409.12191* (2024).

[8] Xu, J., Xiong, Z. and Bhattacharyya, S.: PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers, *CVPR*, pp. 19529–19539 (2023).

[9] Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J. and Zhao, H.: Depth Anything V2, *arXiv:2406.09414* (2024).

[10] Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H. and Torr, P. H.: LAVT: Language-Aware Vision Transformer for Referring Image Segmentation, *CVPR*, pp. 18155–18165 (2022).

[11] Zhang, Y., Wu, D., Han, W. and Dong, X.: Bootstrapping Referring Multi-Object Tracking, *arXiv preprint arXiv:2406.05039* (2024).