Vol.2021-NL-248 No.1
Vol.2021-CVIM-226 No.1
2021/5/20

# 視覚と言語によるナビゲーション課題への言語に対応付けられた生成的な方策

栗田修平[1,a)]　Kyunghyun Cho[2,b)]

**概要**：人間の持っている日常的な知覚 - 視覚や動作の認識，言語理解など - を統合することと，人間から与えられた言語指示に従って判断し行動することは，本来は別の課題である．しかし，近年の視覚と言語による課題を解くニューラルネットワークでは，これらをいずれも入力情報として取り扱い，相互アテンションにより統合しようとしている．このアプローチは既存手法で遍く見られるが，この欠点を指摘するとともに，このアプローチを取らないニューラルネットワークの視覚と言語によるナビゲーション課題への応用を紹介する．

# Generative Language-Grounded Policy in Vision-and-Language Navigation

***Abstract:*** The two abilities of humans: integrating human perceptions – vision, action recognition and language comprehension – and decision-making of actions following the given instructions, are essentially two different tasks. However, recent neural networks for vision-and-language tasks take both of them as input information and try to integrate them by the cross-attention between them. Even though this formalism is frequently applied in existing studies, we will discuss the shortcomings of this formalism and introduce the application of neural networks that do not take this formalism to the vision-and-language navigation tasks.

## 1. はじめに

人間の持っている日常的な知覚を理解し，人間から与えられた指示に従って判断し行動するシステムの構築には，人間の知覚する環境に近い条件でモデル学習を行なっていくことが自然だろう．このようなアイディアは古くから存在した[*1]が，近年，実世界を模した仮想の環境を利用し，仮想のモデルエージェントが環境内部の移動などのインタラクションを通してモデル学習を行うことが現実となりつつある．このような人工知能分野は，*Embodied AI* (身体化された AI; 具体化された AI) と呼ばれ，急速なデータセットや環境整備 [4], [6], [7], [14], [19], [20], [21], [27], [30] とともに活発なモデル提案 [8], [9], [18], [24], [25], [31] が行われ，現実世界への応用 [3], [10] も試みられ始めている．この一つである視覚と言語によるナビゲーション (vision-and-language navigation; VLN) 課題 [4] は，図 1 のように，内部を移動できる写実的な仮想の 3D 屋内環境を利用し，与えられた自然言語による指示に従って仮想モデルエージェントに未知の屋内環境を探索させ，言語指示文章に示された通りに目的の場所まで移動するものである．

視覚と言語によるナビゲーションが目指すところは，かなり現実に近い環境で人間の指示を理解し従うシステムの作成である．このようなシステムでは言語指示に従って動作する課題と，言語と視覚や動作との対応付け課題の双方が求められる．この 2 つは異なる課題である [29] が，既存手法では，指示に従って動作選択を行うことを目標としており，言語と視覚や動作との対応付けについては，相互アテンションなどで取り扱われるのみで明確ではなかった．提案手法では，まず視覚や動作情報を，条件付言語モデル[*2]を用いて明示的に言語情報に写像し，写像を利用して指示文章に従い次の動作を選択する．加えて，提案手法による言語モデルの予測スコアを利用することで，視覚や動

[1] 理化学研究所 革新知能統合研究センター, JST さきがけ
[2] New York University Courant Institute, New York University Center for Data Science, CIFAR Fellow
[a)] shuhei.kurita@riken.jp
[b)] kyunghyun.cho@nyu.edu

[*1] T. Winograd の SHRDLU など [28].

[*2] 本論文の言語モデルは，例えば GPT のような言語生成モデルのことを指し，BERT のようなマスク型言語モデルではない．近年，この 2 つの用法に混乱が見られるため注記する．

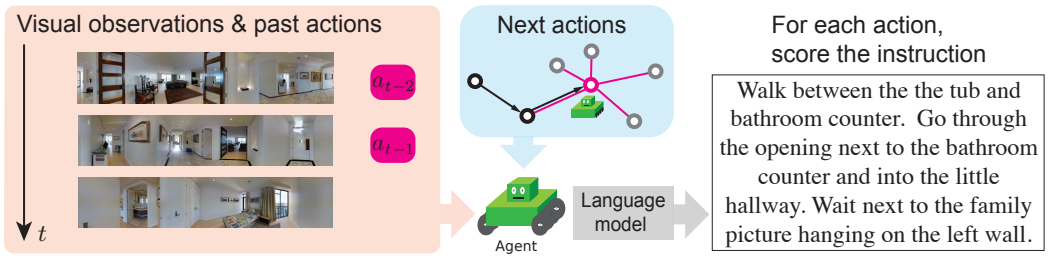


**図 1** 視覚と言語によるナビゲーション課題および言語に対応付けられた生成的な方策．時系列的で写実的な視覚情報と動作情報を，条件付き言語モデルを用いて，最初に与えられた動作指示へと対応付け，次の動作を予測をする．

作情報を元にどのような言語情報に着目してモデルが判断を行なったのか，可視化することが可能となった．

## 2. 視覚と言語によるナビゲーション

視覚と言語によるナビゲーション (VLN) は，内部を移動でき場所に応じた視覚情報が得られる仮想環境を利用して，与えられた言語指示に示された方法で仮想のモデルエージェントを目的の位置まで移動する課題である [4]．広く利用されているデータセットとして，現実の家屋内の 3D スキャンである Matterport3D [5] による写実的な視覚情報を利用した **R2R** [4] と **R4R** [14] が存在する．R2R は，3D スキャン内部にモデルエージェントが移動できるグラフを作成し，エージェントの動作開始地点から課題の目的位置までの最短経路に対して，人手によるエージェントへの経路指示文章がアノテーションされている．R2R はモデルが学習時に未知であった建物環境への汎化能力を測定することを目的としており，Matterport3D の 90 棟分の建物屋内スキャンを，学習データセット，未知 (unseen) 検定データセット，未知テストデータセットへと 61 : 11 : 18 の割合で分割する．それぞれに経路と言語指示アノテーションが作成され，例えば学習データは 14,025 個の経路アノテーションを含む．また，学習データと建物スキャンを共有する既知 (seen) 検定データセットが用意されている．

### 2.1 VLN の問題定式

VLN はモデルが未知の建物の内部を移動して目的位置を探すタスクを目標としている．したがって，探索対象の建物の全体の情報は，タスクの開始時点では一切使用しない．VLN の実験設定でエージェントが利用可能である情報は，初期時刻にて与えられる目的位置までの経路指示文章と，自分が現在いるもしくは過去に訪れた地点の視覚情報，およびそれらの地点において次に進める方向である．

VLN のナビゲーション試行は，反復的な意思決定問題となる．まず課題試行の開始前に，仮想環境中のモデルエージェントに現在位置から目的位置までの経路指示文章 $X$ を与える．ある時刻 $t \in \mathbb{N}$ に，モデルにはその位置に応じて視覚情報 $s_t$[*3]が与えられ，次の動作 $a_t$ を予測する．ただし，各時刻において過去の観測情報も用いることができるため，モデルの入力として用いることができる観測情報を $h_t = \{s_{:t}, a_{:t-1}\}$ としたとき，次の動作 $a_t \in \mathcal{A}_t$ を，例えば確率モデルを用いて $p(a_t|X, h_t)$ と選択する．なお，関連研究 [8], [18], [22] にあるように，可能な動作の集合 $\mathcal{A}_t$ を動作選択において用い，各動作の分散表現を視覚情報から計算し入力として使用できる．可能な動作集合 $\mathcal{A}_t$ には必ず試行終了動作 “STOP” が含まれており，これを選択した地点をモデルエージェントの予測した目的位置とし，ナビゲーション試行の終了とする．このタスク終了時の位置や途中経路をナビゲーション課題の評価に使用する．

また，学習セットの建物環境で学習したモデルを一般の環境に汎化すること，および，テスト時に試行対象となる建物の情報を学習時に使用しないために，VLN では，先述のように Matterport3D の 90 棟分の建物屋内スキャンそのものを，学習データ，未知 (unseen) 検定データ，未知テストデータへと分割している．このような VLN の性質は，学習済みのモデルを現実のロボットへとデプロイすることを可能としている [3]．

### 2.2 VLN の評価手法と R4R

VLN タスクの評価方法として，まず単純に，エージェントが試行終了時に目的位置に存在したか否かを判定する方法がある．この場合，目的とされる位置から 3m 以内で “STOP” 動作を選択した場合に，タスク成功とし，タスク成功割合を成功率 **success rate (SR)** で表す．なお，仮に目的位置から 3m 以内を通り過ぎたとしても，その位置で “STOP” 動作を行わなかった場合には，タスク成功とはならない．$N$ 回の試行に対し $i$ 番目の試行のタスク成否を $S_i \in \{0, 1\}$ とおくと，$SR = \frac{1}{N}\sum_i S_i$ となる．

R2R では必ず最短経路に沿うように経路指示文章が作成されているが，これに対し，エージェントは文章中に指示された経路指示に従っているのか？という指摘が存在した [2]．この問題に対処するために，課題の成功率 **SR** [4] だけではなく，最短経路を基準とし長い経路への罰則を含

[*3] VLN では一般にパノラマ画像を用いる [8].

む **SPL** [2] が考案された．SPL は $S_i$ およびエージェントがナビゲーションに要した経路長 $p_i$，タスク開始地点から終了地点への最短経路長 $l_i$ を用いて

$$SPL = \frac{1}{N}\sum_i S_i \frac{l_i}{\max(p_i, l_i)}$$

と定義される．SPL は，個別の試行に対してタスク失敗時には 0 となり，成功時では最大で 1 を取る．加えて，タスク成功時にも遠回りの経路をたどった場合には値が低下する．SPL は VLN だけではなく，幅広い Embodied AI のタスク評価に利用されている．しかし，SPL は最短距離でのナビゲーション以外には定義されていないことに注意する必要がある．

タスク成功に要した距離だけに着目する SPL に加えて，アノテートされた経路とモデルの経路との近さを測る **CLS** [14] や **nDTW**，**SDTW** [13] も考案されている．CLS や SDTW は，課題の成功率 SR に加えて，エージェントがどのくらい言語指示に従って行動したか，を示す指標である．また，R2R より長く必ずしも最短ではない経路への指示に従う R4R データセットが作られた [14]．R4R では経路指示は最短経路に従わないために，CLS, SDTW を SR に加えて使用する．

VLN では，自分の位置情報を視覚情報から読み取る必要がある一方で，目的の場所や経路は言語指示に従う必要がある．しかし，VLN においても，既存モデルが言語もしくは視覚いずれかの情報に頼りすぎているのではないか，という指摘が存在した [12], [23]．本研究では，このような問題に対処するため，まず視覚および動作を入力とする言語モデルを利用して指示情報を予測し，次に言語指示に従った動作選択を行うモデルを提案する．

### 2.3 VLN の現実世界への応用

また，R2R データセットを利用して機械学習されたモデルを，現実世界の中を移動できるロボット上にデプロイし，ロボットを動かすことが可能である．関連研究 [3] では，R2R における SoTA モデルであった EnvDrop[22] を，家庭用掃除ロボットに類似する小型のロボットに搭載させ，パノラマ画像に相当する 3D カメラを搭載することで，現実の建物内で動作実験を行っている．機械学習による VLN モデルの構築と現実世界でのロボット等の動作との関係性は，この論文の本旨から外れるため付録にて議論する．

### 2.4 言語に対応付けられた生成的な方策

言語および視覚情報によるナビゲーション課題では，以下のように学習パラメータ $\theta$ を最適化する：

$$\max_\theta \sum_t \log p(a_t|h_t, X), \tag{1}$$

既存手法では，この式における $p(a_t|X, h_t)$ を直接モデリングし，深層学習により学習していた [4], [8], [15], [17], [18], [22], [24].

提案手法では，まず動作情報および視覚情報から言語情報へと条件付き言語モデルを利用する写像変換を行い，写像変換の結果を利用して動作選択を行うモデルを提案した．本研究では，まず $p(a_t|X, h_t)$ に対し以下の式変形を行った．

$$\begin{aligned} p(a_t|h_t, X) &= \frac{p(X|a_t, h_t)p'(a_t|h_t)}{\sum_{a'_t\in\mathcal{A}_t} p(X|a'_t, h_t)p'(a'_t|h_t)} \\ &= \frac{p(X|a_t, h_t)}{\sum_{a'_t\in\mathcal{A}_t} p(X|a'_t, h_t)}, \end{aligned} \tag{2}$$

ここで，ある時刻 $t$ で可能な動作集合 $\mathcal{A}_t$ に対し，$p'(a_t|h_t) = 1/|\mathcal{A}_t|$ という近似を用いた．これは仮想エージェントについて，言語指示がなければどのような動作を取る確率も一定であるものと解釈できる．そして，視覚および動作情報から指示文章を予測する，ある種の条件付き言語モデル $p(X|a'_t, h_t)$ を深層学習を用いてモデルし，式 2 に従って動作予測する手法を提案した．既存手法のアプローチは，動作の条件付き確率を予測する点で識別的 (Discriminative) な方策であり，提案手法はベイズ的な意味で生成的 (Generative) な方策である．

最適化は以下のように行われる．

$$\max_\theta \sum_t \Big(\log p(X|a_t, h_t) - \log \sum_{a'_t\in\mathcal{A}_t} p(X|a'_t, h_t)\Big). \tag{3}$$

この第一項 $\log p(X|a_t, h_t)$ は正解動作 $a_t$ が与えられた際の通常の言語モデルの学習の式と同一であるが，第二項は，正解動作 $a_t$ を含む可能なすべての動作に対する罰則項である．本研究では，$p(X|a_t, h_t)$ について，条件付言語モデルとして学習済みのモデルを使用し，式 3 により VLN のための学習を行う．

このような定式化を採用する利点は，視覚や動作情報を条件付言語モデルを利用して言語情報と明確に対応付け，その情報を利用して指示に従う動作を選択することである．既存手法のように $p(a_t|X, h_t)$ を直接モデリングする場合，視覚情報と言語情報の対応付けには広く相互アテンションが用いられてきた．しかし，相互アテンションは視覚と言語をどのように対応付けて課題を解いているか明確でなく，片方を無視して予測を行うことも起こりうる[*4]．提案手法では，条件付き言語モデル $p(X|a_t, h_t)$ をモデルに取り入れることで，まずは視覚情報及び動作情報を指示文章へ明確に結びつける必要がある．3.4 節では，このような条件付言語モデルによる指示文章の予測スコアが，VLN の探索のために学習済みのモデル内部でどのように振る舞うか検証する．

[*4] アテンションは説明として妥当か長い議論がある [26].

| | Validation (Seen) | | | | | | | Validation (Unseen) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | PL↓ | NE↓ | SR↑ | SPL↑ | CLS↑ | nDTW↑ | SDTW↑ | PL↓ | NE↓ | SR↑ | SPL↑ | CLS↑ | nDTW↑ | SDTW↑ |
| Disc. | 10.69 | 5.40 | 0.519 | 0.482 | 0.619 | 0.588 | 0.445 | 12.88 | 6.52 | 0.380 | 0.335 | 0.488 | 0.458 | 0.304 |
| Disc. +Aug.(A) | 10.60 | 5.15 | 0.525 | 0.489 | 0.633 | 0.596 | 0.445 | 12.05 | 6.22 | 0.431 | 0.392 | 0.528 | 0.496 | 0.356 |
| Gen. | 11.23 | 5.53 | 0.481 | 0.451 | 0.625 | 0.579 | 0.427 | 12.98 | 6.17 | 0.434 | 0.371 | 0.514 | 0.478 | 0.344 |
| Gen. +Aug. (B) | 11.45 | 4.78 | **0.563** | **0.531** | **0.664** | **0.630** | **0.505** | 13.92 | 4.78 | **0.476** | **0.405** | **0.539** | **0.503** | **0.379** |
| Gen.+Disc.(A+B) | 10.18 | 4.67 | 0.568 | 0.540 | 0.680 | 0.640 | 0.510 | 12.06 | 5.42 | 0.489 | 0.437 | 0.570 | 0.533 | 0.403 |
| Gen.+Disc.(A+B)* | 11.30 | 4.58 | 0.575 | 0.541 | 0.678 | 0.636 | 0.509 | 14.65 | 5.19 | 0.518 | 0.439 | 0.564 | 0.515 | 0.397 |

**表 1** 生成的な方策 (Gen.) と識別的な方策 (Disc.) の R2R 検定データセットによる性能比較結果．+Aug. は関連研究 [8] にて提案された拡張データの使用を示す．(A) Disc.+Aug. と (B) Gen.+Aug. を識別的および生成的な手法をあわせて用いる実験に使用する (A+B)．∗ は back-tracking[15] の使用を示す．太字は単一モデルとして最良の性能を表す．

| | Validation (Seen) | | | | Validation (Unseen) | | | | Test (Unseen) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | PL↓ | NE↓ | SR↑ | SPL↑ | PL↓ | NE↓ | SR↑ | SPL↑ | PL↓ | NE↓ | SR↑ | SPL↑ |
| Seq2seq [4] | 11.33 | 6.01 | 0.39 | - | 8.39 | 7.81 | 0.22 | - | 8.13 | 7.85 | 0.20 | 0.18 |
| Speaker-Follower [8] | - | 3.36 | 0.66 | - | - | 6.62 | 0.35 | - | 14.82 | 6.62 | 0.35 | 0.28 |
| EnvDrop [22] | 11.0 | 3.99 | 0.62 | 0.59 | 10.70 | 5.22 | 0.52 | 0.48 | 11.66 | 5.23 | 0.51 | **0.47** |
| FAST* [15] | - | - | - | - | 21.17 | 4.97 | 0.56 | 0.43 | 22.08 | 5.14 | **0.54** | 0.41 |
| PRESS [17] | 10.57 | 4.39 | 0.58 | 0.55 | 10.36 | 5.28 | 0.49 | 0.45 | 10.77 | 5.49 | 0.49 | 0.45 |
| Gen.+Disc. Policy | 10.18 | 4.67 | 0.57 | 0.54 | 12.06 | 5.42 | 0.49 | 0.44 | 11.90 | 5.52 | 0.51 | **0.46** |
| Gen.+Disc. Policy* | 11.30 | 4.58 | 0.57 | 0.54 | 14.65 | 5.19 | 0.52 | 0.44 | 14.31 | 5.24 | **0.54** | **0.46** |
| Human | - | - | - | - | - | - | - | - | 11.90 | 1.61 | 0.86 | 0.76 |

**表 2** R2R テストセットにおける比較．太字は SR と SPL が一番目，二番目に良いモデルを示す．

## 3. 実験

### 3.1 実験設定・モデル

VLN にて広く使われる R2R データセットと R4R データセットにて，生成的な方策と既存手法による識別的な方策のいずれが優れているか，比較実験を行った．実験には，関連研究 D. Fried らの Speaker-follower モデル [8] と同一構造を持つニューラルネットワークを用いた．Speaker-follower モデルは，視覚情報と言語指示を入力としナビゲーション課題を解く Follower モデルと，動作情報および動作に従った際の視覚情報から言語指示文章を生成する Speaker モデルからなる．特に Speaker モデルには単独で R2R ナビゲーション課題を解く能力はないが，D. Fried らは，Speaker モデルを R2R データセットのデータ拡張 (data augmentaton) および，ビーム探索を利用したナビゲーション結果の再ランキングに Follower と合わせて使用した．なお，R2R ナビゲーション課題でビーム探索を使用することは，現実世界では同時に複数台のロボットを使用して目的位置を探索することに相当し，あまり意味のある実験設定とは言えないため，本論文では考察しない．実際に，VLN のタスク提案者を含むグループは，VLN にてビーム探索を使用しないように求めている [2]．

本研究では，この D. Fried らの Speaker モデルを言語に対応付けられる生成的な方策に用いる．画像情報の符号化には ResNet-152 [11] を使用する．また，D. Fried らの Follower モデルを，既存手法にて典型的に用いられる識別的な方策として同様に学習に用いる．学習には PRESS [17] と同様の学習を用いる[*5]．また，FAST [15] にて提案された，モデルエージェントが以前に自分が訪れた場所まで行動罰則付きで戻る事ができる back-tracking をモデルの評価時に採用した場合の結果を示す．Fried らによって公開されているデータ拡張された学習データが広くモデル学習に採用されていることを考慮し，多くの関連研究 [8], [15], [17] と同様にこの拡張データを学習に使用する[*6]．実験では，先に紹介した主な評価尺度である SR, SPL, CLS, nDTW, SDTW に加えて，エージェントの平均探索距離 **PL**，タスク終了時の目的地への平均距離 **NE** を補助的に示す．

なお，言語モデルそのものである D. Fried らの Speaker

[*5] なお，PRESS は BERT を使用しており，本研究と同じく事前学習済みモデルを使用するモデルである．
[*6] EnvDrop [22] ではさらに多くの拡張データを使用している．

| Model | Validation (Seen) PL↓ | NE↓ | SR↑ | CLS↑ | nDTW↑ | SDTW↑ | Validation (Unseen) PL↓ | NE↓ | SR↑ | CLS↑ | nDTW↑ | SDTW↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RCM fidelity-oriented[13] | 18.8 | 5.4 | 0.526 | 0.553 | - | - | 28.5 | 5.4 | 0.261 | 0.346 | - | - |
| nDTW fidelity-oriented[13] | - | - | - | - | - | - | - | - | 0.285 | 0.354 | 0.304 | 0.126 |
| BabyWalk IL+RL[31] | - | - | - | - | - | - | 22.8 | 8.6 | 0.250 | 0.455 | 0.344 | 0.136 |
| BabyWalk IL+RL+Cur.[31] | - | - | - | - | - | - | 23.8 | 7.9 | 0.296 | **0.478** | **0.381** | **0.181** |
| Disc. supervised | 20.1 | 7.0 | 0.386 | 0.622 | 0.512 | 0.305 | 20.0 | 9.8 | 0.172 | 0.446 | 0.305 | 0.101 |
| Disc. fidelity-oriented | 21.1 | 6.6 | 0.449 | 0.644 | 0.530 | 0.360 | 29.2 | 9.2 | 0.211 | 0.385 | 0.282 | 0.116 |
| Gen. supervised | 19.8 | 8.8 | 0.316 | 0.563 | 0.442 | 0.246 | 19.7 | 9.8 | 0.193 | **0.479** | 0.325 | 0.121 |
| Gen. fidelity-oriented | 21.0 | 6.9 | 0.448 | 0.629 | 0.517 | 0.349 | 22.8 | 8.7 | 0.255 | 0.471 | **0.348** | **0.162** |

**表 3** 生成的な方策 (Gen.) と識別的な方策 (Disc.) の R4R データセットにおける性能比較．太字は未知データセットにて CLS，nDTW，SDTW が一番目，二番目に良いモデルを示す．

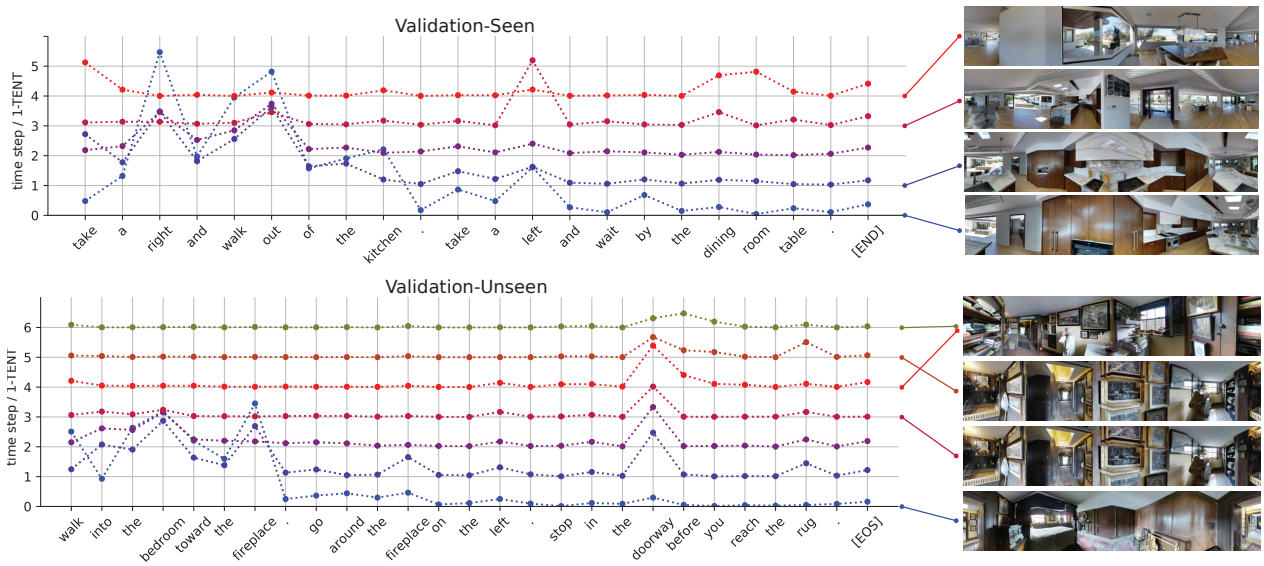


**図 2** トークンごとの予測エントロピー ($1 - TENT$)．既知検定データセット（上段）と未知検定データセット（下段）での例．．縦軸は時刻 $t$ と $1 - TENT$ の振幅を表す．横軸は指示文章中のトークンであり，既知検定データセットでは “Take a right and walk out of the kitchen. Take a left and wait by the dining table.” に対応する．

モデルに VLN を解く能力は存在しないが，言語モデル事前学習を行わずに式 3 の学習を行った場合は，精度が大きく低下する．なお，式 1 に従ってマスク型言語モデルである事前学習済み BERT を指示文章の符号化に利用する研究?が存在するが，著者の再現実験では，BERT の効果は特に未知データセットにて明確ではなかった．

### 3.2 R2R データセットでの実験結果

まず，R2R における生成的な方策と既存手法において典型的な識別的方策との比較を，R2R の検定データセットで行った．結果を表 1 に示す．データ拡張済みの学習データでは生成的な方策は識別的な方策を常に上回ることがわかる．また，生成的な方策と識別的な方策をモデル評価時に結合したモデル (Gen.+Disc) はそれらをさらに上回るとともに，back-tracking [15] の使用により改善が見られる．

次にテストセットで既存手法との比較を行った．結果を表 2 に示す．提案手法は，R2R 評価において広く使われる SR および SPL の両方の評価尺度にて，先行研究と同等以上の性能を達成している．

また，R2R よりも長距離の経路指示に従う R4R でも実験を行った．結果の詳細は付録に示すが，概略を述べると，R4R の未知セットの性能にて生成的な方策は識別的な方策を常に上回り，提案手法は CLS にて先行研究と同等以上の性能を達成した．

### 3.3 R4R による実験結果

表 3 は R4R における実験結果を示す．R4R では，経路上での単純な教師あり学習 (supervised) と，言語指示された経路に沿う形での学習 (fidelity-oriented) の 2 つのモデル学習を，識別的 (Disc.) および生成的 (Gen.) な方策の双方に対し行った．なお，R4R データセットにはテストセットが存在しないため，既知 (seen) 検定セットと未知 (unseen) 検定セットでの結果を示す．R4R の既存手法では SR が高いわりに CLS や SDTW が低いものも存在し，これは指示文書にて示された経路に従わずに目的位置を探索している可能性を示唆する．提案手法は一般に未知データセットにおけるスコアが高く，学習環境に過適合しにくい可能性がある．

### 3.4 言語に対応付けられた説明性

本論文では，言語モデルに与えられる動作入力に応じて，言語モデルの予測がどのように変化するかを可視化するため，トークンごとの予測エントロピー ($TENT$) を提案した．$TENT$ は，あるトークンの予測スコアが動作の集合の中でどのくらい変化するかエントロピーを用いて導出しており，$1-TENT$ が大きいトークンほど，動作予測に大きな影響を持つと考えられる．

指示文章 $X = [w_1, \cdot, w_k, \cdot]$ に対し，$TENT = S(w_k)$ は以下のように定義され，$S(w_k) \in [0, 1]$ を満たす．

$$S(w_k) = -\sum_{a_t \in \mathcal{A}} q(a_t, w_k) \log_{|\mathcal{A}|} q(a_t, w_k), \tag{4}$$

$$q(a_t, w_k) = \frac{p(w_k|a_t, h_t, w_{:k-1})}{\sum_{a_t \in \mathcal{A}_t} p(w_k|a_t, h_t, w_{:k-1})}. \tag{5}$$

上式で，$\sum_{a_t \in \mathcal{A}_t}$ は，ある時刻 $t$ での可能な動作集合 $\mathcal{A}_t$ 要素全てに対する和であることを思い起こすと，このトークンごとの予測エントロピーは，そのトークン予測スコアが動作入力に応じてどのように変化するかを示す指標となる．さらに，可視化の都合上，$1-TENT$ を考えると，$1-TENT$ が大きいトークンほど，動作予測に大きな役割を持つと考えられる．

なお，下式ではトークン予測スコアの動作集合による正規化を行っている．これは，トークン毎の予測スコアにはそもそも言語構造により差が存在し，その効果を平均操作により打ち消し，動作の違いによるエントロピーを式 4 で評価するためにある．例えば，命令文 “take a left”(左に曲がれ) において，“a” の予測確率と “left” の予測確率はそもそも異なると考えられるが，このようなトークン毎の予測スコアの変化は，動作予測とは無関係であると思われるために，この正規化を行っている．. 図 2 は，既知および未知検定セットのデータ例について，$1-TENT$ を図示したものである．動作の開始時 ($t \sim 0$) には，トークン予測が安定しない性質がある（グラフ振幅が大きい）ものの，各時刻において，予測の要となりやすいトークンが視覚化されており，例えば，$t = 3$ では，“take a left” の “left” という単語予測スコアが，動作選択の決め手になったことを示唆している．

**図 3** ALFRED [21] の動作環境である AI2THOR [16] 環境において，エージェントの動作計画が難しい例．エージェントの主観視点画像から動作計画を行う．エージェントは左前方のベッドと右のソファの間を通り抜けて，右前方へと通り抜けたいとする．現地点で一度でも前進操作を行うと，エージェントはベッドフレームに衝突する．右旋回，前進，左旋回，前進と行うことで通り抜けることが出来るが，右旋回の後，2 回にわたって前進操作を行うと，ソファーに衝突する．

## 4. 議論

### 4.1 sim2real 問題: VLN エージェントを実世界のロボットにデプロイできるか

VLN により学習されたモデルは，現実世界のロボットにデプロイすることが目的されており，実際に Mattreport 仮想環境で学習されたデータを利用して現実世界のロボットを動かすことが可能である．関連研究 [3] は，R2R のように予め経路グラフが作成されアノテーションされた状態で，仮想環境上で 55.9%という SR に対し，現実世界で 46.8%という SR を達成している．さらに驚くべきことに，事前に収集されたマップ情報がない場合でも，22.5%という SR を達成している．マップ情報が存在しない場合，関連研究 [3] では，自己の近傍位置からエージェントの次の移動可能な方向を予測する subgoal モジュールおよび SLAM[*7] と同時に機械学習モデルである VLN エージェントを使用することで，移動可能な方向を導出し，実際の移動を可能にしている．また，Matterport3D 環境と，実際にロボットを動かすための ROS シミュレータ上での精度が，EnvDrop モデル [22] における実験で 1%以内で一致したという．

[*7] 自己位置推定と環境地図作成

### 4.2 Embodied AI における個別動作の実行可能性の判定についての検討

VLN においてロボットの可能な動作集合 $\mathcal{A}$ は予測の対象外にある．これはある状況においてロボットが取ることができる動作は，少なくとも VLN モデルによる機械学習モデルの判定外にあることを意味している．これは，先行研究である [8] や [18] においても同様の実験設定を取っている．すなわち，VLN エージェントは，自分が選択した動作が不可能であることは想定していない．なお，当然のことながら，ある動作が実際に実行することができなかった場合に，その動作を外して再度，動作予測を行うことは可能である．

VLN と同様に Embodied AI 環境でのタスクを解く ALFRED [21] では，エージェントがある時点で可能な動作を，動作の判定前に知ることはできない．したがって，ALFRED では，機械学習モデルが次の動作予測としてエージェントが実行不可能な動作を予測することが起こり得る．ALFRED の仮想環境での試行時には，10 回の無効な動作選択を行った際はタスク終了としている．このような性質は，学習をひどく困難にしている可能性がある．例えば，図 3 の例では，深度センサーのような物理情報を使用せずに，動作計画を立てることは難しいだろう．

逆に Habitat シミュレータ [19] を用いた実験では，エージェントが壁に斜めにぶつかった場合に，壁に沿った方向に向かってスライドしてしまう問題が知られている．壁に沿ったスライドが許されている場合，エージェントは通路の中央を通り抜けるような動作計画を行うことがなく目的位置までたどり着くことが可能となるが，実際のロボットに紐付いた動作とは言えないだろう．なお，このスライディングを禁止した Habitat Challenge 2020 [1] では，モデルの動作精度 (SPL) が，Habitat Challenge 2019 と比較して大きく低下している．

このように，ある時点でロボットが可能な動作の集合 $\mathcal{A}$ は，現段階では，ある程度，VLN のように動作計画を行う機械学習モデルの外側から与えられるべきであると考える．これは，ある動作がそのロボットにとって適か不適かは，ロボットの物理的なセンサなどから，その動作を行う前に安全性などの観点に基づいて判断されるのが適切と考えるからである．実際に，このように学習されたモデルを，ロボット側のモデルと組み合わせることで実世界におけるナビゲーションが可能とされる [3]．同時に，特に言語指示を実際の動作計画へと結びつける問題設定においては，可能な動作集合から選択を繰り返した結果，目的状態への動作計経路が袋小路に陥らないように，ある程度の先の状態の予測を行いながら個別の動作選択を行う必要があるだろう．

## 5. 結論

視覚と言語によるナビゲーションは，人間の知覚環境に近い仮想環境上で機械学習を行う革新的な課題であるが，しかし，言語と視覚やその他の情報がどのように対応付けられているか，既存手法では明らかではない側面があった．提案手法のアプローチが，このように視覚と言語，動作など複数のモダリティを結合するシステムを作成するための重要な示唆となることを願ってやまない．

**謝辞** 本研究は JST ACT-I，JPMJPR17U8 および JST さきがけ，JPMJPR20C2 の支援を受けました．NYU 訪問につき関根聡先生および関係する皆様に感謝します．なお，この論文は言語処理学会第 27 回年次大会で発表済みの論文「視覚と言語によるナビゲーション課題への言語に対応付けられた生成的な方策」を合同研究会向けにまとめ直し内容を追加したものです．

### 参考文献

[1] Abhishek Kadian*, Joanne Truong*, Gokaslan, A., Clegg, A., Wijmans, E., Lee, S., Savva, M., Chernova, S. and Batra, D.: Are We Making Real Progress in Simulated Environments? Measuring the Sim2Real Gap in Embodied Visual Navigation, *arXiv:1912.06321* (2019).

[2] Anderson, P., Chang, A. X., Chaplot, D. S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M. and Zamir, A. R.: On Evaluation of Embodied Navigation Agents, *ArXiv*, Vol. abs/1807.06757 (2018).

[3] Anderson, P., Shrivastava, A., Truong, J., Majumdar, A., Parikh, D., Batra, D. and Lee, S.: Sim-to-Real Transfer for Vision-and-Language Navigation, *CoRL* (2020).

[4] Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S. and van den Hengel, A.: Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments, *Proceedings of the IEEE CVPR* (2018).

[5] Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A. and Zhang, Y.: Matterport3D: Learning from RGB-D Data in Indoor Environments, *International Conference on 3D Vision (3DV)* (2017).

[6] Chen, H., Suhr, A., Misra, D., Snavely, N. and Artzi, Y.: TOUCHDOWN: Natural Language Navigation and Spatial Reasoning in Visual Street Environments, *Proceedings of the IEEE/CVF CVPR* (2019).

[7] Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D. and Batra, D.: Embodied Question Answering, *Proceedings of the IEEE CVPR* (2018).

[8] Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.-P., Berg-Kirkpatrick, T., Saenko, K., Klein, D. and Darrell, T.: Speaker-Follower Models for Vision-and-Language Navigation, *Advances in Neural Information Processing Systems 31*, Curran Associates, Inc., pp. 3314–3325 (online), available from ⟨http://papers.nips.cc/paper/7592-speaker-follower-models-for-vision-and-language-navigation.pdf⟩ (2018).

[9] Hao, W., Li, C., Li, X., Carin, L. and Gao, J.: Towards Learning a Generic Agent for Vision-and-Language

Navigation via Pre-training, Vol. abs/2002.10638, (online), available from ⟨https://arxiv.org/abs/2002.10638⟩ (2020).

[10] Hatori, J., Kikuchi, Y., Kobayashi, S., Takahashi, K., Tsuboi, Y., Unno, Y., Ko, W. and Tan, J.: Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions, *Proceedings of ICRA* (2018).

[11] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, IEEE Computer Society, pp. 770–778 (online), DOI: 10.1109/CVPR.2016.90 (2016).

[12] Hu, R., Fried, D., Rohrbach, A., Klein, D., Darrell, T. and Saenko, K.: Are You Looking? Grounding to Multiple Modalities in Vision-and-Language Navigation, *Proceedings of the 57th Annual Meeting of the ACL*, Florence, Italy, pp. 6551–6557 (online), DOI: 10.18653/v1/P19-1655 (2019).

[13] Ilharco, G., Jain, V., Ku, A., Ie, E. and Baldridge, J.: General Evaluation for Instruction Conditioned Navigation using Dynamic Time Warping, *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop, Vancouver, Canada, December 13, 2019*, (online), available from ⟨https://vigilworkshop.github.io/static/papers/33.pdf⟩ (2019).

[14] Jain, V., Magalhaes, G., Ku, A., Vaswani, A., Ie, E. and Baldridge, J.: Stay on the Path: Instruction Fidelity in Vision-and-Language Navigation, *Proceedings of the 57th Annual Meeting of the ACL*, Florence, Italy, pp. 1862–1872 (online), DOI: 10.18653/v1/P19-1181 (2019).

[15] Ke, L., Li, X., Bisk, Y., Holtzman, A., Gan, Z., Liu, J., Gao, J., Choi, Y. and Srinivasa, S.: Tactical Rewind: Self-Correction via Backtracking in Vision-And-Language Navigation, *The IEEE CVPR* (2019).

[16] Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A. and Farhadi, A.: AI2-THOR: An Interactive 3D Environment for Visual AI, *arXiv* (2017).

[17] Li, X., Li, C., Xia, Q., Bisk, Y., Celikyilmaz, A., Gao, J., Smith, N. A. and Choi, Y.: Robust Navigation with Language Pretraining and Stochastic Sampling, *Proceedings of the 2019 Conference on EMNLP-IJCNLP*, Hong Kong, China, pp. 1494–1499 (online), DOI: 10.18653/v1/D19-1159 (2019).

[18] Ma, C., Lu, J., Wu, Z., AlRegib, G., Kira, Z., Socher, R. and Xiong, C.: Self-Monitoring Navigation Agent via Auxiliary Progress Estimation, *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, (online), available from ⟨https://openreview.net/forum?id=r1GAsjC5Fm⟩ (2019).

[19] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D. and Batra, D.: Habitat: A Platform for Embodied AI Research, *Proceedings of the IEEE/CVF ICCV* (2019).

[20] Nguyen, K. and Daumé III, H.: Help, Anna! Visual Navigation with Natural Multimodal Assistance via Retrospective Curiosity-Encouraging Imitation Learning, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 684–695 (online), DOI: 10.18653/v1/D19-1063 (2019).

[21] Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L. and Fox, D.: ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks, *The IEEE CVPR*, (online), available from ⟨https://arxiv.org/abs/1912.01734⟩ (2020).

[22] Tan, H., Yu, L. and Bansal, M.: Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout, *Proceedings of the 2019 Conference of the NAACL, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, pp. 2610–2621 (online), DOI: 10.18653/v1/N19-1268 (2019).

[23] Thomason, J., Gordon, D. and Bisk, Y.: Shifting the Baseline: Single Modality Performance on Visual Navigation & QA, *Proceedings of the 2019 Conference of the NAACL, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, pp. 1977–1983 (online), DOI: 10.18653/v1/N19-1197 (2019).

[24] Wang, X., Huang, Q., Çelikyilmaz, A., Gao, J., Shen, D., Wang, Y., Wang, W. Y. and Zhang, L.: Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation, *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, Computer Vision Foundation / IEEE, pp. 6629–6638 (online), DOI: 10.1109/CVPR.2019.00679 (2019).

[25] Wang, X., Xiong, W., Wang, H. and Wang, W. Y.: Look Before You Leap: Bridging Model-Free and Model-Based Reinforcement Learning for Planned-Ahead Vision-and-Language Navigation, *ECCV*, (online), available from ⟨https://eccv18-vlease.github.io/static/papers/look-before-you-leap.pdf⟩ (2018).

[26] Wiegreffe, S. and Pinter, Y.: Attention is not not Explanation, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Association for Computational Linguistics, pp. 11–20 (online), DOI: 10.18653/v1/D19-1002 (2019).

[27] Wijmans, E., Datta, S., Maksymets, O., Das, A., Gkioxari, G., Lee, S., Essa, I., Parikh, D. and Batra, D.: Embodied Question Answering in Photorealistic Environments with Point Cloud Perception, *Proceedings of the IEEE CVPR* (2019).

[28] Winograd, T.: Procedures as a Representation for Data in a Computer Program for Understanding Natural Language (1971).

[29] Wittgenstein, L.: Philosophische Untersuchungen (1953).

[30] Wu, Y., Wu, Y., Gkioxari, G. and Tian, Y.: Building generalizable agents with a realistic and rich 3D environment, *arXiv preprint arXiv:1801.02209* (2018).

[31] Zhu, W., Hu, H., Chen, J., Deng, Z., Jain, V., Ie, E. and Sha, F.: BabyWalk: Going Farther in Vision-and-Language Navigation by Taking Baby Steps, *CoRR*, Vol. abs/2005.04625 (online), available from ⟨https://arxiv.org/abs/2005.04625⟩ (2020).