AIGeN: An Adversarial Approach for Instruction Generation in VLN

Niyati Rawal   Roberto Bigazzi   Lorenzo Baraldi   Rita Cucchiara

University of Modena and Reggio Emilia

{name.surname}@unimore.it

Abstract

In the last few years, the research interest in Vision-and-Language Navigation (VLN) has grown significantly. VLN is a challenging task that involves an agent following human instructions and navigating in a previously unknown environment to reach a specified goal. Recent work in literature focuses on different ways to augment the available datasets of instructions for improving navigation performance by exploiting synthetic training data. In this work, we propose AIGeN, a novel architecture inspired by Generative Adversarial Networks (GANs) that produces meaningful and well-formed synthetic instructions to improve navigation agents' performance. The model is composed of a Transformer decoder (GPT-2) and a Transformer encoder (BERT). During the training phase, the decoder generates sentences for a sequence of images describing the agent's path to a particular point while the encoder discriminates between real and fake instructions. Experimentally, we evaluate the quality of the generated instructions and perform extensive ablation studies. Additionally, we generate synthetic instructions for 217K trajectories using AIGeN on Habitat-Matterport 3D Dataset (HM3D) and show an improvement in the performance of an off-the-shelf VLN method. The validation analysis of our proposal is conducted on REVERIE and R2R and highlights the promising aspects of our proposal, achieving state-of-the-art performance.

1. Introduction

In the last decade, we have witnessed remarkable results in Natural Language Processing, Computer Vision, and Robotics [6]. More recently, increasing interest has been devoted to research at the intersection between these three domains [5, 31, 44]. In line with this trend, our work focuses on the Vision-and-Language Navigation (VLN) task. When performing VLN, an agent or a robot can perceive the 360° view of the environment and is given human instructions such as "*Go to the living room and bring me the remote on the table*". The agent has to follow the instructions and navigate an unknown environment to reach the specified goal and stop there.

However, human-generated instructions needed for training such architectures are costly and time-consuming to obtain. The resulting lack of annotated training data is one of the key factors in making VLN a challenging task. Recent work [11, 16, 20, 56] has focused on generating instructions at a lower cost by employing methods for synthetic instruction generation. For example, Guhur *et al.* [20] and Chen *et al.* [11] showed that generating synthetic instructions and augmenting the data, improves the navigation performance of the agent. Nevertheless, Guhur *et al.* [20] used image-caption pairs from the web on a prohibitive total number of 140K environments. Instead, Chen *et al.* [11] generated synthetic instructions using trajectories sampled on HM3D dataset [47] which is composed of 900 environments; our approach aims at improving the generation quality of such instructions.

As can be seen in Fig. 1, we propose AIGeN, a novel computational model that can generate synthetic instructions starting from unlabeled navigation paths in an environment. Our proposed model combines a multimodal Generative Pre-Trained Transformer (GPT) [7] and Bidirectional Encoder Representations from Transformers (BERT) [15] in an adversarial manner to generate high-quality instructions. In particular, the model consists of a Transformer decoder (GPT-2) that generates sentences describing the agent's path, using a sequence of images from the environment and the associated object detections extracted using Mask2Former [13]. The BERT-like encoder, instead, serves as a discriminator and is trained to distinguish between real and fake instructions.

Using our approach to augment the training data of REVERIE and R2R datasets, we show that our AIGeN-generated instructions help to improve the results of a VLN model achieving state-of-the-art performance. Additionally, we validate the quality of the proposed method by evaluating the generated instructions using image description metrics [49] and comparing the downstream navigation performance of different configurations of our model, showing that producing well-formed synthetic instructions is beneficial for the training of a navigation agent.
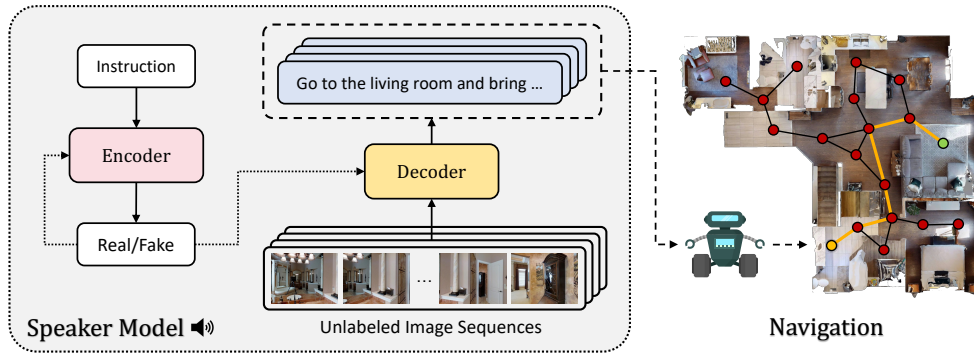
Figure 1. We present AIGeN, a novel GAN-like model for generating instructions given a sequence of images. Synthetic instructions can be used as training data for a VLN model to improve its navigation performance.

2. Related Work

Vision-and-Language Navigation. VLN is a task where an agent needs to learn to navigate in a previously unknown environment by following step-by-step natural language instructions [3, 19]. Thus, the VLN task requires the agent to understand visual and textual cues and their correlation and plan its movements accordingly. This makes VLN a challenging task, which has attracted remarkable research efforts. While this work focuses on VLN in indoor environment, there are also works that emphasize VLN in outdoor environments [9, 42, 58] Early work proposed a sequence-to-sequence long short-term memory for action inference [3] and adopted a panoramic action space with a module to generate synthetic instructions [16]. Ma *et al.* [39] proposed a self-monitoring agent, and a method combining progress inference with learned heuristic [40]. Landi *et al.* [29] used dynamic convolution filters. RCM [57], instead, used reinforcement learning to improve cross-modal matching. Recently, Transformer-based models [52] have been gaining popularity among VLN methods. For example, PRESS [33] applied pretrained BERT to encode instructions. VLN↺BERT [24] implemented a recurrent BERT to model time dependencies. SIA [37] and PTA [30] used Transformers for multimodal fusion, and HAMT [10] directly used Transformers to exploit episode history. Among graph-based methods, Hong *et al.* [23] implemented graphs to model relations between the scene, objects, and instructions, while DUET [12] and KERM [34] used topological maps with a dual-scale graph Transformer to encode long-term action planning and fine-grained cross-modal understanding. Additionally, AZHP [17] proposed a hierarchical navigation process instead of a single-step planning scheme.

Instruction Generation. A different direction of research on VLN is followed by studies that present methods for augmenting available datasets. Besides the already mentioned Fried *et al.* [16], Majumdar *et al.* [41] proposed a Trans-

former model that transfers visual grounding learned from image-text pairs from the web to VLN. Other research work samples random trajectories and generates instructions using online rental marketplaces [20] and a large-scale dataset of indoor environments [11, 27]. Our approach improves the generation of synthetic instructions by aligning the visual input features with the generated text using the same number of trajectories as Chen *et al.* [11].

The most common Transformer-like networks used in Natural Language Processing (NLP) applications are BERT [15], GPT [7], BART [32] and Text-to-Text Transfer Transformer (T5) [46]. While BERT consists of only the encoder part of the Transformer, GPT consists of its decoder part. BERT is beneficial for tasks like Visual Question Answering (VQA), *e.g.* [38, 50]. On the other hand, GPT is useful for text generation or summarization [7]. Moreover, Visual Language Models (VLMs) [1] have been used to generate text, given images, and textual words. Chen *et al.* [11] proposed to use a GPT-2 decoder that acts like a speaker model to generate instructions for VLN. In this study, we use a GPT-2 decoder that generates language instructions, while BERT encoder discriminates whether the instructions correspond to the given sequence of images as real or fake.

Generative Adversarial Networks. While GANs have been applied effectively for generating images [18], they have been seldom applied to generate instructions that consist of discrete tokens as it is difficult to backpropagate through discrete tokens. Previous works rely on REINFORCE [14] or using Gumbel-Softmax [48] to allow for backpropagation. Recently, Transformer networks were used in an adversarial manner, however, these approaches were used for the generation of images [25, 26].

3. Proposed Method

AIGeN is used to describe an agent's trajectory in natural language and is composed of an instruction generator (visually depicted in Fig. 2) and an instruction discriminator.
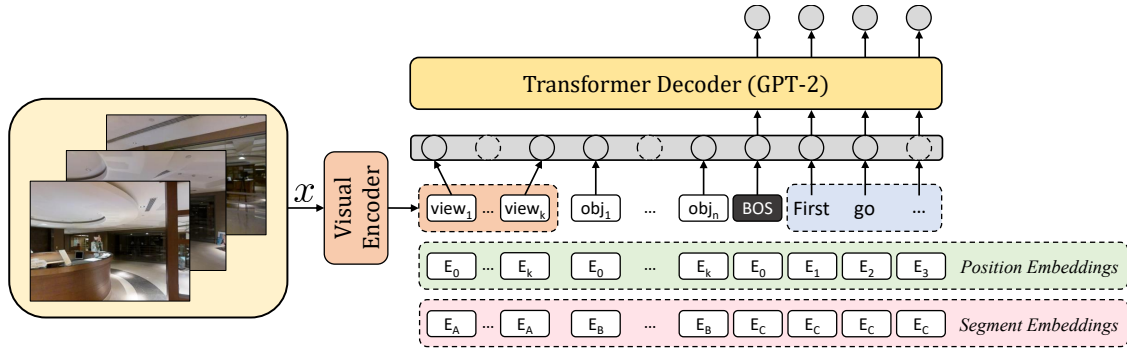
Figure 2. Architecture of the instruction generator model that generates VLN instructions. The model is fed with a sequence of images and object names, and it is prompted to generate instructions using a `BOS` token. The instructions are generated token by token until the `EOS` token.

Both models are trained simultaneously using an adversarial approach, which aims at improving the performance of the generator and the quality of the generated instructions. The overall model is shown in Fig. 3. To the best of our knowledge, this is the first approach that combines Transformer networks and a GAN-like training procedure for generating synthetic navigation instructions.

3.1. Instruction Generator

The generator is a language model in charge of describing the actions needed to reach the target location. The proposed model exploits a pretrained language model which is finetuned conditioning on visual inputs to achieve multimodal capabilities similar to Alayrac *et al.* [1]. However, instead of just using GPT-2 decoder for instruction generation, we use it in combination with BERT in a GAN-like manner to generate better-quality instructions.

The overall approach takes the images from the environment the agent has to traverse and feeds them into a pretrained ResNet-152 [22] encoder to obtain image features. Next, when generating instructions on Habitat-Matterport 3D (HM3D) environments, all the objects in the given sequence of images are extracted using Mask2Former [13] trained on ADE20K to enrich the visual features with object names. This input is fed into the decoder along with the first `BOS` token using the concatenation of visual features and object names as a prompt for the language model. The GPT-2 decoder is trained to predict the subsequent language tokens to generate a complete language instruction that describes the actions of the agent until the last image. Differently from [1], AIGeN flattens the tokens from both images and text before feeding them into the GPT-2 decoder. Furthermore, to effectively segregate the visual information from the textual information, position and segment embeddings are used in addition to the tokens, as shown in Fig. 2.

The result of the generation is defined starting from the output of the decoder following the `BOS`, autoregressively generating instructions token by token and allowing back-

propagation through sampled tokens using the Gumbel-Softmax [48] trick.

3.2. Instruction Discriminator

The inputs of the BERT encoder are image features, tokens of the names of objects in the images, and an instruction. Similarly to the inputs for GPT-2 decoder, the images are first fed into a pretrained ResNet-152 [22] visual encoder to obtain image features. Then, a fully connected layer followed by a sigmoid function is used to process the results from BERT output. The output is simply real or fake; real if the given instruction matches the given sequence of images and fake otherwise. Finally, binary cross-entropy loss is used to minimize the error between the generated output and the actual output (real or fake). A well-trained discriminator should classify ground-truth instructions as real and the generated instructions as fake. Like the decoder, BERT input employs position embeddings and segment embeddings in addition to token embeddings.

3.3. Adversarial Training

Our method, shown in Fig. 3, follows an adversarial training approach, where the generator G is trained to fool the discriminator D, while the discriminator is taught to distinguish between real and fake instructions. The generator is trained to generate instructions as close to the ground-truth instructions as possible by minimizing the cross-entropy loss between the generated instructions and the ground-truth instructions. The generator loss is defined as:

$$\mathcal{L}_G = -\log(D(I_G, \boldsymbol{x})), \qquad (1)$$

where $I_G \in G(\boldsymbol{x})$ is the generated instruction and $\boldsymbol{x}$ is the sequence of images belonging to the trajectory.

The discriminator has to discriminate between the ground-truth instructions as real instructions and the generated instructions as fake instructions. Consequently, the discrimi-
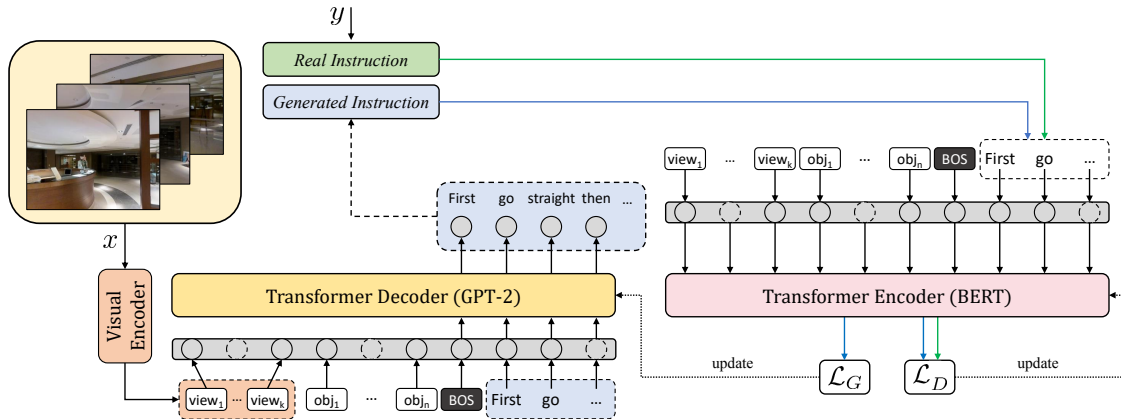
Figure 3. Schema of the proposed generative-adversarial framework for synthetic instruction generation. The GPT-2 decoder acts as a generator while the BERT encoder acts as a discriminator. Both models are trained simultaneously. The generator generates fake instructions token by token until it reaches the EOS token. The discriminator has to detect whether the instructions corresponding to a given sequence of images is real (ground-truth) or fake (generated).

nator loss is:

$$\mathcal{L}_D = -\log(1 - D(I_G, \boldsymbol{x})) - \log(D(I_R, \boldsymbol{x})), \quad (2)$$

where $I_R \in R(\boldsymbol{x})$ is the ground-truth instruction. The training is performed simultaneously both on the generator and the discriminator.

4. Experiments

4.1. Experimental Setup

We present experimental results using the model proposed in Section 3 employing two VLN datasets as benchmark, REVERIE [45] and R2R [3]. Both datasets consist of navigation sequences composed of 360° images that are collected on the nodes of navigation graphs on Matterport3D environments [8] and each navigation sequence is associated with three ground-truth textual instructions. The main difference between the two datasets is that REVERIE instructions include interactions with specific target objects that the model is required to identify, whereas R2R instructions only specify the trajectory to be followed. In this study, we only consider the frontal view of the 360° images, with a field of view of 60°. The training of AIGeN uses a learning rate of 0.0002, batch size of 6, and Adam [28] as optimizer. For the GPT-2 decoder, unless otherwise stated we use a medium-sized version with $L = 12$, $d = 768$, $H = 12$, where L is the number of layers, d is the model dimensionality and H is the number of attention heads. The visual features used by the model are extracted using ResNet-152. For the BERT encoder, we use a hidden size of 768, 12 layers, and an intermediate size of 3072. Overall, AIGeN has 289M model parameters and is trained for ≈ 36 hours on a single NVIDIA RTX6000 GPU.

We evaluate the navigation performance using the following navigation metrics: trajectory length in meters (TL); success rate (SR), *i.e.* the fraction of episodes where the agent can reach the goal position within 3 meters; oracle success rate (OSR), that is the success rate using an oracle stop policy; success rate weighted by path length (SPL); and navigation error (NE). The object grounding ability of the agents on REVERIE dataset is evaluated using remote grounding success (RGS) which is the ratio of successfully followed instructions, and RGS weighted by path length (RGSPL).

Instead, to quantitatively evaluate our model on text generation, we use metrics commonly used for image description [49], namely BLEU [43], METEOR [4], ROUGE [36], CIDEr [53], and SPICE [2][1]. All these metrics are obtained by comparing the predicted instruction with the ground-truth instruction in terms of their *n*-grams (where an *n*-gram is a sequence of *n* consecutive words). While all these metrics are commonly used for evaluating cross-modal description, only CIDEr and SPICE have been specifically designed for this task. The others (BLEU, METEOR, and ROUGE) have indeed been proposed for evaluating translation and summarization. According to recent literature, CIDEr showcases the best alignment with human judgment.

4.2. Experimental Results on VLN

The navigation experiments on VLN are performed considering an off-the-shelf state-of-the-art VLN method finetuned with the instructions generated by our approach. We adopt DUET [12], which is based on a dual-scale graph Transformer to perform long-term action planning and fine-grained cross-modal understanding by exploiting topological maps that are built during the episode. DUET has a two-step training as described by Chen *et al.* [12], that is

---

[1]We compute image description metrics using the code provided at the following link: https://github.com/tylin/coco-caption

	Val Unseen					
	TL	SPL↑	SR↑	OSR↑	RGS↑	RGSPL↑
Seq2Seq [3]	11.1	2.8	4.2	8.1	2.2	1.6
SMNA [39]	9.1	6.4	8.2	11.3	4.5	3.6
RCM [57]	12.0	7.0	9.3	14.2	4.9	3.9
FAST-MATTN [45]	45.3	7.2	14.4	28.2	7.8	4.7
SIA [37]	41.5	16.3	31.5	44.7	22.4	11.6
Airbert [20]	18.7	21.9	27.9	34.5	18.2	14.2
ProbES [35]	18.0	22.8	27.6	33.2	16.8	13.9
VLN↻BERT [24]	16.8	24.9	30.7	35.0	18.8	15.3
HAMT [10]	14.1	30.2	33.0	36.8	18.9	17.3
DUET [12]	22.1	33.7	47.0	51.1	32.2	23.0
KERM [34]	21.9	35.4	50.4	55.2	34.5	24.5
AZHP [17]	22.3	36.6	48.3	53.6	34.0	25.8
HM3D-AutoVLN* [11]	24.3	39.4	54.0	**60.0**	34.6	25.2
DUET + AIGeN	19.5	**41.9**	54.4	57.7	**35.1**	**26.9**

Table 1. VLN metrics for our approach and competitors on REVERIE Val Unseen. * denotes finetuning on our computing architecture.

	Val Unseen				**Test Unseen**			
	TL	SPL↑	SR↑	NE↓	TL	SPL↑	SR↑	NE↓
Seq2Seq [3]	8.39	-	22	7.81	8.13	18	20	7.85
PRESS [33]	10.36	45	49	5.28	10.77	45	49	5.49
SSM [54]	20.70	45	62	4.32	20.39	46	61	4.57
EnvDrop [51]	10.70	48	52	5.22	11.66	47	51	5.23
PREVALENT [21]	10.19	53	58	4.71	10.51	51	54	5.30
RelGraph [23]	9.99	53	57	4.73	10.29	52	55	4.75
ProbES [35]	11.58	55	61	4.03	12.43	56	62	4.20
Airbert [20]	11.78	56	62	4.01	12.41	57	62	4.13
VLN↻BERT [24]	12.01	57	63	3.93	12.35	57	63	4.09
MARVAL [27]	10.15	61	65	4.06	10.22	58	62	4.18
DUET [12]	13.94	60	72	3.31	14.73	59	69	3.65
KERM [34]	13.54	60	72	3.22	14.74	59	70	3.61
HAMT [10]	11.46	61	66	**2.29**	12.27	60	65	3.93
AZHP [17]	14.05	61	72	3.15	14.95	60	**71**	3.52
DUET + AIGeN	13.72	**63**	**73**	2.92	14.20	**61**	71	**3.33**

Table 2. VLN experiments on the Val Unseen split of R2R dataset.

composed of a pretraining phase on four auxiliary tasks and the finetuning on the VLN task. We perform the pretraining phase following [11], and then we finetune the pre-trained model augmenting REVERIE and R2R training splits with our synthetic instructions. AIGeN-generated instructions are produced using 217K randomly sampled trajectories on Habitat-Matterport 3D (HM3D) [47] dataset, a large-scale dataset of indoor photorealistic environments. We use the same paths used and released by Chen *et al.* [11], and we generate one synthetic instruction for each sequence of observations. The navigation model is pre-trained and finetuned for a total of $\approx$ 32 hours, 12 and 20 respectively, on a single NVIDIA RTX6000 GPU, and is compared with current state-of-the-art VLN models on REVERIE and R2R.

The navigation results on REVERIE in Table 1 show that our approach achieves state-of-the-art performance on SPL and both object grounding metrics while remaining competitive on the other metrics. In particular, our instruction set gives a boost in generating effective trajectories towards the goal, which is reflected by SPL and RGSPL. For example, our approach shows an improvement of 8.2 and 3.9 on SPL and RGSPL with respect to the baseline DUET. For fairness of comparison, we re-trained HM3D-AutoVLN on our computing infrastructure following [11] (denoted as HM3D-AutoVLN*). The comparison between HM3D-AutoVLN* and our approach highlights the effectiveness of AIGeN-generated instruction since the difference between the two methods on VLN is defined by the quality of the synthetic instructions exploited to finetuned DUET. When running this comparison, our approach outperforms HM3D-AutoVLN* on the main navigation metrics, and in particular, on success rate weighted by path length (SPL) shows an improvement of 2.5.

Furthermore, analyzing the experiments on R2R dataset in Tab. 2, our method provides state-of-the-art results on SPL and SR for both validation and test unseen splits, prov-

	Val Seen					Val Unseen				
Model	BLEU-1	METEOR	ROUGE	CIDEr	SPICE	BLEU-1	METEOR	ROUGE	CIDEr	SPICE
AIGeN (Medium)	**0.487**	0.222	0.457	0.869	0.318	0.412	0.166	0.378	0.461	0.213
AIGeN w/o detect.	0.409	0.188	0.416	0.358	0.196	0.357	0.141	0.347	0.132	0.123
AIGeN	0.484	**0.228**	**0.465**	**0.890**	**0.329**	**0.421**	**0.179**	**0.393**	**0.486**	**0.228**

Table 3. Image description experiments using our model with different size configurations and without using the detections on REVERIE.

	Val Unseen					
	TL	SPL↑	SR↑	OSR↑	RGS↑	RGSPL↑
AIGeN (Medium)	20.0	41.2	54.3	**60.1**	**35.1**	26.5
AIGeN w/o detect.	23.4	38.9	52.5	59.4	34.6	25.7
AIGeN	19.6	**41.9**	**54.4**	57.7	**35.1**	**26.9**

Table 4. Experimental comparison of VLN performance of different configurations of our model on REVERIE dataset.

	%Novel	Unigrams	Bigrams	Div-1	Div-2
Ground truth	-	3675	21551	0.019	0.113
AIGeN w/o adv. training	22.7%	2970	15928	0.016	0.086
AIGeN	100.0%	14783	43013	0.072	0.210

Table 5. Evaluation of the diversity of synthetic instructions using AIGeN before and after the GAN fine-tuning compared with the ground truth instructions on REVERIE training split.

ing that our approach can generate synthetic instructions that are beneficial for multiple VLN datasets. Comparing our method with respect to the baseline DUET, AIGeN-generated instructions allow an improvement in SPL of 3.0 and 2.0 respectively on the validation "unseen" and test "unseen" splits. Overall, these results support our claim that well-formed synthetic instructions help the agent to learn better navigation and object localization in a VLN setting.

4.3. Ablation Study

We validate the components of our approach by comparing different configurations of AIGeN on synthetic instructions generation in Table 3 and on navigation in Table 4.

Starting from Table 3, the first row shows the performance of the generator that is trained from scratch. As can be seen, training the instruction generation model from scratch provides a maximum CIDEr of 0.869 on the "seen" split, and 0.378 on the "unseen" split, without reaching the results of the overall method. Following, we ablate the proposed model in terms of input modalities analyzing the contribution given by the object detections, and comparing it with a model that is trained using only visual features and textual instructions. All the metrics related to AIGeN without detections, especially CIDEr and SPICE, are considerably lower than the metric values computed for AIGeN. This result confirms the importance of employing object detections as input features. We speculate that when there are no object words provided, the model is not able to identify which object in the scene it has to attend to. Therefore, the target objects in the generated sentences are often different

from those in the ground-truth instructions even when the landmark is correctly recognized.

Moving on to the navigation experiments in Table 4, we compare the performance of the VLN model on REVERIE dataset augmenting the training data with the instructions generated by different configurations of our approach. In this case, AIGeN surpasses its counterparts by an important margin on all the metrics. The improvement on the SPL over the Medium configuration that is trained from scratch is 0.7, and it becomes 3.0 when considering AIGeN w/o detections. The ablation study validates the effectiveness of the components of the model in both text generation and downstream vision-and-language navigation.

4.4. Diversity Analysis

To assess the quality of the synthetic instructions and measure the diversity of the generation, in Table 5 we compute diversity metrics commonly used for image captioning [49, 55] comparing the diversities of AIGeN-generated instructions with the ground-truth annotations of the trajectories of REVERIE training split. The comparison is performed by producing synthetic instruction using AIGeN before and after the adversarial fine-tuning phase.

The metrics used for this study are the ratio of the number of novel sentences, *i.e.* not contained in the dataset, to the number of ground-truth sentences, the number of unique words (Unigrams); the number of unique couples of consecutive words (Bigrams); Div-1; and Div-2. Div-1 and Div-2 are respectively the ratio of unique unigrams and bigrams to the total number of unigrams.

(a) **GT**: *Go to the white room with 4 basins opposite the lounge pool and turn on the faucet on the left closest to the entrance*
AIGeN: *Walk to the bathroom with the white sink and turn on the faucet*

(b) **GT**: *Go to the hallway with the black chairs on level 2 and sit in the second chair from your right*
AIGeN: *Go to the dining room and bring me the chair that is nearest to the door*

(c) **GT**: *Go to the lounge on level 1 and sit down on the armchair*
AIGeN: *Go to the dining room and with a lint roller clean the arm chair which is against the wall*

(d) **GT**: *Go to the laundry room on the first floor and sort out the cluttered cleaning products on the dryer*
AIGeN: *Go to the laundry room on level 1 and bring me the clothes on the rack opposite the dryer*

Figure 4. Sample image sequences from REVERIE Val Unseen split with corresponding ground-truth instruction and synthetic instructions generated using AIGeN. The images in each sequence have been reduced to 8 to facilitate the graphical presentation and we only show the frontal image of the panoramic observation at each timestep.

Looking at the results, it is evident that the fine-tuning phase using an adversarial approach helps to generate instructions that do not retrace the ground truth instruction, with the consequence of improving the diversity of the dataset. In fact, while the generated instructions without using the GAN-like training present a small number of novel sentences, AIGeN returns a completely novel set of instructions. Furthermore, AIGeN with the adversarial fine-tuning can increase the number of unigrams and bigrams sampled from the word dictionary even with respect to the ground truth annotations. This result is also reflected by Div-1 and Div-2 metrics that present a significant increase with respect to the ground truth instructions.

4.5. Qualitative Analysis

Finally, Fig. 4 and 5 show examples of sequences of input images with corresponding ground-truth instructions (if available) and generated instructions using AIGeN.

In Fig. 4 all four samples have been taken from the "unseen" validation split of REVERIE, so that AIGeN has never seen these environments during training. We provide two positive samples in (a) and (b) as well as two nega-

tive ones in (c) and (d). For both (a) and (b), the generated instruction is similar to the ground-truth instruction and matches the given sequence of images. In fact, the landmarks bathroom in (a) and the dining room in the hallway in (b) are correctly recognized. Furthermore, both synthetic instructions refer to the correct target object, respectively "faucet" and "chair". In the case of (c) instead, the generated instruction and the ground-truth instruction identify the wrong landmark, with the lounge that is recognized as a dining room by our model. However, the target object "armchair" is still recognized correctly. Finally, in the case of (d), the generated instruction and the ground-truth one, refer to different target objects. While the correct target objects are the "cluttered cleaning products" on the dryer, AIGeN refers to the "clothes on the rack" as the target objects. Nevertheless, the laundry room is correctly identified by our model.

Moving on to Fig. 5, we showcase three qualitative samples using trajectories from HM3D. As HM3D is unlabeled, in this case, there is no ground-truth annotation available, and we only provide the instructions generated by AIGeN. The first example in (e) identifies the office and the target
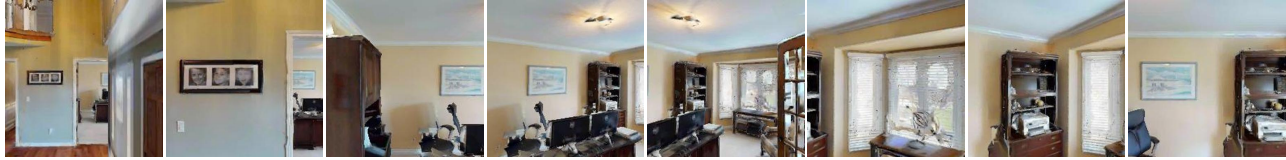
(e) **AIGeN**: *Go to the office on level 1 and clean the desk*

(f) **AIGeN**: *Go to the kitchen and turn off the light closest to the entrance*

(g) **AIGeN**: *Go to the laundry room and clean the mirror above the sink*

Figure 5. Sample image sequences from HM3D with corresponding synthetic instructions generated using AIGeN. The images in each sequence have been reduced to 8 to facilitate graphical presentation, and we only show the frontal image of the panoramic observation at each timestep.

object "desk" correctly in the final observation of the sequence. In the second trajectory shown in (f), the target object ("light") in the kitchen is correctly recognized. Finally, in the third trajectory (g), the kitchen is identified as a laundry room, and the correct target object "window" is misidentified as a "mirror".

These results demonstrate that even when the visual quality of the environment is low due to 3D reconstruction, AIGeN is capable of generating valuable instructions for vision-and-language navigation providing correct directions identifying objects and landmarks.

5. Conclusions

We proposed a computational model for synthetic instruction generation, based on the state-of-the-art NLP models GPT-2 and BERT. Our model leverages a GAN-like structure and takes sequences of images as input to generate instructions for the path traversed by the agent to reach its target location. The model has been trained and validated on REVERIE instructions and achieved high image description metric results when comparing the generated instructions with the ground-truth instructions. Subsequently, we prove that using our synthetic instructions to augment a VLN dataset, such as REVERIE or R2R, improves the performance of a VLN method achieving state-of-the-art performance on both navigation and object grounding metrics.

6. Limitations

In our experimental section, we extract the labels of the objects in the last panoramic view of each navigation trajectory using Mask2Former [13] trained on ADE20K. The labels of these objects are tokenized to obtain their embeddings which are then concatenated with the image embeddings and the instructions embeddings. Relying on detection labels the model can identify the target objects within each episode and potentially understand the surrounding context and landmarks. This time-consuming procedure is necessary as the words derived from the object labels help to generate coherent instructions for the provided image sets.

As can be seen in the ablation study presented in Table 3 and Table 4, the instruction quality and navigation performance of AIGeN without the object detections are worse than those of the complete approach. Future work could focus on generating instructions directly from the image sequence trying to fill the gap with the approach using upstream object detection.

Acknowledgments

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems*, 2022. 2, 3

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. In *Proceedings of the European Conference on Computer Vision*, 2016. 4

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 2, 4, 5

[4] Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshops*, 2005. 4

[5] Roberto Bigazzi, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. Embodied Agents for Efficient Exploration and Smart Scene Description. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2023. 1

[6] Roberto Bigazzi, Lorenzo Baraldi, Shreyas Kousik, Rita Cucchiara, and Marco Pavone. Mapping High-level Semantic Regions in Indoor Environments without Object Recognition. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2024. 1

[7] 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. In *Advances in Neural Information Processing Systems*, 2020. 1, 2

[8] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *Proceedings of the International Conference on 3D Vision*, 2017. 4

[9] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural Language Navigation and Spatial Reasoning in Visual Street Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[10] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History Aware Multimodal Transformer for Vision-and-Language Navigation. *Advances in Neural Information Processing Systems*, 2021. 2, 5

[11] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Learning from Unlabeled 3D Environments for Vision-and-Language Navigation. In *Proceedings of the European Conference on Computer Vision*, 2022. 1, 2, 5

[12] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think Global, Act Local: Dual-scale Graph Transformer for Vision-and-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 4, 5

[13] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3, 8

[14] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards Diverse and Natural Image Descriptions via a Conditional GAN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017. 2

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018. 1, 2

[16] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-Follower Models for Vision-and-Language Navigation. *Advances in Neural Information Processing Systems*, 2018. 1, 2

[17] Chen Gao, Xingyu Peng, Mi Yan, He Wang, Lirong Yang, Haibing Ren, Hongsheng Li, and Si Liu. Adaptive Zone-aware Hierarchical Planner for Vision-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 5

[18] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, 2014. 2

[19] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022. 2

[20] Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. Airbert: In-Domain Pretraining for Vision-and-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 5

[21] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards Learning a Generic Agent for Vision-and-Language Navigation via Pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 5

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 3

[23] Yicong Hong, Cristian Rodriguez, Yuankai Qi, Qi Wu, and Stephen Gould. Language and Visual Entity Relationship

Graph for Agent Navigation. *Advances in Neural Information Processing Systems*, 2020. 2, 5

[24] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. VLN BERT: A Recurrent Vision-and-Language BERT for Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 5

[25] Drew A Hudson and Larry Zitnick. Generative Adversarial Transformers. In *Proceedings of the International Conference on Machine Learning*, 2021. 2

[26] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Trans-GAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up. *Advances in Neural Information Processing Systems*, 2021. 2

[27] Aishwarya Kamath, Peter Anderson, Su Wang, Jing Yu Koh, Alexander Ku, Austin Waters, Yinfei Yang, Jason Baldridge, and Zarana Parekh. A New Path: Scaling Vision-and-Language Navigation with Synthetic Instructions and Imitation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 5

[28] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *Proceedings of the International Conference on Learning Representations*, 2014. 4

[29] Federico Landi, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. Embodied Vision-and-Language Navigation with Dynamic Convolutional Filters. In *Proceedings of the British Machine Vision Conference*, 2019. 2

[30] Federico Landi, Lorenzo Baraldi, Marcella Cornia, Massimiliano Corsini, and Rita Cucchiara. Multimodal Attention Networks for Low-Level Vision-and-Language Navigation. *Computer Vision and Image Understanding*, 2021. 2

[31] Federico Landi, Roberto Bigazzi, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. Spot the Difference: A Novel Task for Embodied Agents in Changing Environments. In *Proceedings of the International Conference on Pattern Recognition*, 2022. 1

[32] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020. 2

[33] Xiujun Li, Chunyuan Li, Qiaolin Xia, Yonatan Bisk, Asli Celikyilmaz, Jianfeng Gao, Noah Smith, and Yejin Choi. Robust Navigation with Language Pretraining and Stochastic Sampling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2019. 2, 5

[34] Xiangyang Li, Zihan Wang, Jiahao Yang, Yaowei Wang, and Shuqiang Jiang. KERM: Knowledge Enhanced Reasoning for Vision-and-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 5

[35] Xiwen Liang, Fengda Zhu, Lingling Li, Hang Xu, and Xiaodan Liang. Visual-Language Navigation Pretraining via Prompt-based Environmental Self-exploration. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022. 5

[36] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshops*, 2004. 4

[37] Xiangru Lin, Guanbin Li, and Yizhou Yu. Scene-Intuitive Agent for Remote Embodied Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 5

[38] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViL-BERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*, 2019. 2

[39] Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-Monitoring Navigation Agent via Auxiliary Progress Estimation. In *Proceedings of the International Conference on Learning Representations*, 2019. 2, 5

[40] Chih-Yao Ma, Zuxuan Wu, Ghassan AlRegib, Caiming Xiong, and Zsolt Kira. The Regretful Agent: Heuristic-Aided Navigation through Progress Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[41] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving Vision-and-Language Navigation with Image-Text Pairs from the Web. In *Proceedings of the European Conference on Computer Vision*, 2020. 2

[42] Harsh Mehta, Yoav Artzi, Jason Baldridge, Eugene Ie, and Piotr Mirowski. Retouchdown: Adding Touchdown to StreetLearn as a Shareable Resource for Language Grounding Tasks in Street View. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, 2020. 2

[43] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002. 4

[44] Samuele Poppi, Niyati Rawal, Roberto Bigazzi, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. Towards Explainable Embodied Navigation and Recounting. In *Proceedings of the International Conference on Image Analysis and Processing*, 2022. 1

[45] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 4, 5

[46] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 2020. 2

[47] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-Matterport 3D Dataset (HM3D): 1000

Large-scale 3D Environments for Embodied AI. *Advances in Neural Information Processing Systems*, 2021. 1, 5

[48] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017. 2, 3

[49] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From Show to Tell: A Survey on Deep Learning-based Image Captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 4, 6

[50] Hao Tan and Mohit Bansal. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2019. 2

[51] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019. 5

[52] 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*, 2017. 2

[53] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015. 4

[54] Hanqing Wang, Wenguan Wang, Wei Liang, Caiming Xiong, and Jianbing Shen. Structured Scene Memory for Vision-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 5

[55] Qingzhong Wang, Jia Wan, and Antoni B Chan. On Diversity in Image Captioning: Metrics and Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 6

[56] Su Wang, Ceslee Montgomery, Jordi Orbay, Vighnesh Birodkar, Aleksandra Faust, Izzeddin Gur, Natasha Jaques, Austin Waters, Jason Baldridge, and Peter Anderson. Less is More: Generating Grounded Navigation Instructions from Landmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1

[57] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2, 5

[58] Wanrong Zhu, Xin Wang, Tsu-Jui Fu, An Yan, Pradyumna Narayana, Kazoo Sone, Sugato Basu, and William Yang Wang. Multimodal Text Style Transfer for Outdoor Vision-and-Language Navigation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, 2021. 2