AIGeN-Llama: An Adversarial Approach for Instruction Generation in VLN using Llama2 Model

Niyati Rawal*, Lorenzo Baraldi and Rita Cucchiara

University of Modena and Reggio Emilia

Abstract

Vision-and-Language Navigation (VLN) aims to train a robot to perceive the surrounding environment and follow human instructions. In the context of Digital Libraries, such agents hold transformative potential for assisting users in navigating large, multi-modal repositories and in interpreting and connecting spatial, visual, and textual data. However, training agents to follow human-like instructions in unknown environments remains a significant challenge, largely due to the scarcity of labeled training data. To address this, we propose AIGeN-Llama, an adversarial framework that utilizes Llama2 models for instruction generation. The Llama2 generator synthesizes navigation instructions by processing image sequences, while a Llama2 discriminator determines the authenticity of these instructions compared to ground-truth data. This adversarial training enhances the realism of the generated instructions. We use metrics that are commonly used for image description, namely BLEU, METEOR, ROUGE, CIDEr, and SPICE to quantitatively evaluate the proposed model. In addition, we show some qualitative samples to prove the effectiveness of our method. The experiment highlights the flexibility and capability of Llama2 as both a generator and a discriminator, demonstrating its potential to advance embodied VLN tasks.

Keywords

vision, language, navigation

1. Introduction

Vision-and-Language Navigation (VLN) represents a critical frontier in embodied AI, where agents are tasked with navigating unfamiliar environments based on natural language instructions. Beyond its traditional applications in assistive robotics and autonomous systems, VLN holds significant promise for enhancing digital libraries by enabling more intuitive, interactive, and accessible ways of exploring complex, multi-modal repositories. For instance, VLN agents could guide users through immersive virtual archives or assist in retrieving spatially or thematically relevant digital content using conversational queries. Currently, the development of robust VLN agents remains hindered by the scarcity of large-scale, high-quality datasets that pair trajectories with human instructions. This limitation not only affects generalization to unseen environments, a core requirement for real-world deployment, but also constrains the potential integration of VLN technologies into innovative digital library applications.

Recent studies have shown that augmenting training datasets with synthetic instructions can improve navigation performance [1, 2, 3]. Despite these advances, generating realistic and contextually grounded instructions remains a challenge. Traditional approaches often rely on architectures, such as GPT-2 and BERT, which may lack the flexibility and expressive power of newer large language models (LLMs). To address this, we introduce AIGeN-Llama, an adversarial framework designed to leverage the advanced generative and discriminative capabilities of Llama2, a state-of-the-art LLM.

AIGeN-Llama builds on the principles of adversarial learning, employing Llama2 as both the instruction generator and discriminator (see Fig. 1 for an overview). The generator produces detailed navigation instructions based on image trajectories, while the discriminator evaluates the authenticity and alignment of these instructions with ground-truth data. This adversarial interplay pushes the

IRCDL 2025: 21st Conference on Information and Research Science Connecting to Digital and Library Science, February 20–21, 2025, Udine, Italy

*Corresponding author.

niyti.rawal@unimore.it (N. Rawal); lorenzo.baraldi@unimore.it (L. Baraldi); rita.cucchiara@unimore.it (R. Cucchiara)

0000-0002-4142-0488 (N. Rawal); 0000-0001-5125-4957 (L. Baraldi); 0000-0002-2239-283X (R. Cucchiara)

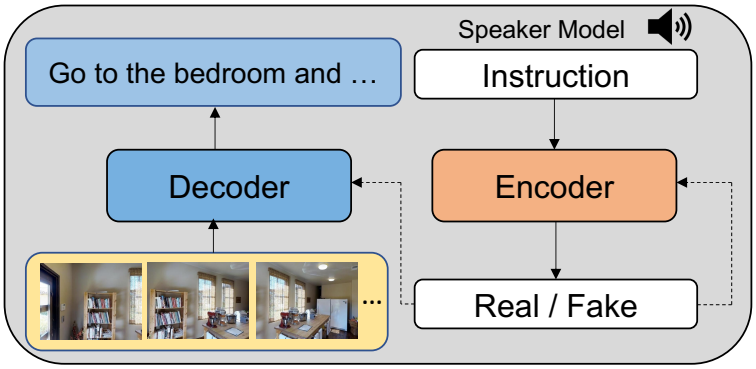


Figure 1: We present the overview of AIGeN-Llama, an adversarial framework that utilizes Llama2 models for instruction generation. AIGeN-Llama consists of a Llama2 encoder and a Llama2 decoder. Llama2 decoder act as a generator and Llama2 encoder act as a discriminator. Both the generator and the decoder are trained simultaneously to generate instructions corresponding to the given sequence of images.

generator to create more realistic and nuanced instructions and also equips the discriminator to refine its ability to distinguish between synthetic and ground-truth instructions.

The motivation for adopting Llama2 lies in its demonstrated ability to excel in a variety of complex generative and understanding tasks, supported by its large-scale pretraining and fine-tuning on diverse datasets. By integrating Llama2 into an adversarial framework, AIGeN-Llama seeks to overcome the limitations of previous architectures, generating more relevant synthetic instructions. To quantitatively evaluate AIGeN-Llama, we use metrics that are commonly used for image description, namely, BLEU, METEOR, ROUGE, CIDEr and SPICE. In addition, we present some qualitative samples that show the ability of AIGeN-Llama to generate reasonable instructions. Our approach sets a new standard in VLN instruction generation and demonstrates the broader applicability of Llama2 in embodied AI systems.

2. Related Work

The field of Vision-and-Language Navigation (VLN) has seen significant advancements in recent years, driven by innovations in both data augmentation and model architectures. AIGeN-Llama builds upon these developments, addressing challenges in synthetic instruction generation and adversarial learning.

2.1. Vision and Language Navigation (VLN)

Vision-and-Language Navigation (VLN) is a challenging task requiring agents to navigate in 3D environments guided by natural language instructions. The Room-to-Room (R2R) dataset by Anderson et al. [4] established a benchmark for VLN, pairing navigation trajectories with human-written instructions. While early works on VLN focused on sequence-to-sequence long short-term memory model for action inference, recent works rely on Transformers [5, 6, 7]. Graph-based methods where graphs are used to model relations between scene, object and instructions [8] or the use of topological maps [9] have also been introduced recently.

2.2. Instruction Generation for VLN

Instruction generation has emerged as a critical task for enhancing VLN datasets. Anderson et al. introduced the Room-to-Room (R2R) dataset, which paired human-authored instructions with trajectories, but highlighted the challenge of scaling such datasets due to the cost of manual annotation [4].

Recent efforts have explored generating synthetic instructions to augment VLN datasets. For instance, Speaker-Follower models [10] synthesized path descriptions but often produced overly simplistic or repetitive instructions. Other research studies generate instructions by sampling random trajectories, leveraging online rental marketplaces [2] and large-scale datasets of indoor environments [1, 11, 3].

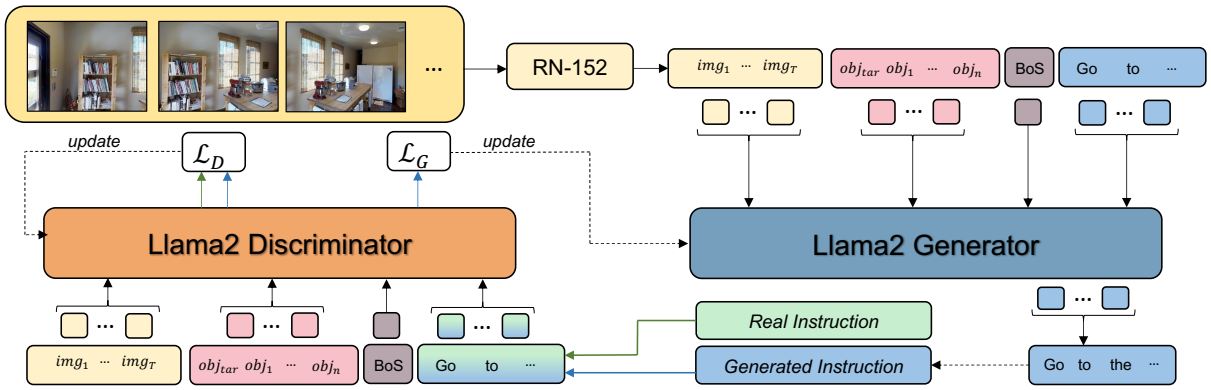


Figure 2: Schema of the proposed generative-adversarial framework for synthetic instruction generation. The Llama2 decoder acts as a generator while the Llama2 encoder acts as a discriminator. The generator generates fake instructions token-by-token until it reaches the EOS token. The discriminator must detect whether the instructions corresponding to a given sequence of images are real (ground truth) or fake (generated).

These methods emphasize the need for high-quality synthetic data to improve the generalization capabilities of navigation agents.

2.3. Large Language Models (LLMs) in VLN

The advent of large-scale pretrained language models, such as GPT and BERT, has had a significant impact on VLN tasks. Recent studies have incorporated GPT-based decoders to generate instructions and BERT-based encoders to contextualize trajectories [3]. However, these models often lack the versatility and power of newer LLMs, such as Llama2, which excel at capturing long-range dependencies and generating more coherent text.

AIGeN-Llama leverages Llama2 for both generative and discriminative roles. Its superior performance in language modeling enables the generation of nuanced and contextually relevant instructions, surpassing prior architectures in quality.

2.4. Adversarial Learning

Adversarial learning, popularized by Generative Adversarial Networks (GANs) [12], has been widely adopted to improve synthetic data generation across various domains, including images, text, and audio. In instruction generation, adversarial learning ensures that generated outputs closely mimic human-like text. Works like [13, 14] demonstrated the potential of adversarial training for text generation To overcome the problem of gradient propagation for discrete outputs, techniques like the Gumbel-Softmax trick [14] were introduced to approximate differentiable sampling. AIGeN-Llama adopts this approach, allowing Llama2 to generate high-quality instructions in an adversarial setting. The discriminator, also powered by Llama2, effectively distinguishes between real and synthetic instructions, pushing the generator toward greater realism and alignment with human-authored data.

3. Method

AIGeN-Llama is an adversarial framework that leverages Llama2 as both a generator and a discriminator to produce realistic and high-quality navigation instructions for VLN. Unlike previous approaches that rely on GPT-2 and BERT, AIGeN-Llama utilizes Llama2's advanced language capabilities to generate more relevant instructions. See Fig. 2 for the schema of the overall model.

3.1. Llama2 Generator

The generator is responsible for creating synthetic instructions based on sequences of images that represent navigation trajectories. It processes the input visual data and sequentially generates tokens, crafting instructions in natural language that guide the agent along the given trajectory.

The general approach is as follows. First, the images of the trajectory are fed into a pretrained ResNet-152 to extract the visual features. Next, all objects in the last image of the trajectory are detected using Mask2Former [15] trained on ADE20K. This is essential to enrich the visual representation. The visual features along with the object names are fed into the Llama2 decoder as input. This is followed by the BOS token which is used by the model as an indication to start generating the instruction for the given trajectory. The Llama2 decoder is trained to predict the next token and predicts autoregressively until it reaches the EOS token. Formally,

$$y = \texttt{Llama2}\left(\left[v_0, .., v_{t_{\text{Images}}}, o_{tgt}, o_0.., o_{n,\text{Objects}}\text{BOS}, i_1, .., i_{m,\text{Instruction}}\text{EOS}\right]\right) \tag{1}$$

where $(v_0, ..., v_t)$ denotes the set of visual features for images of the trajectory, o_{tgt} indicates the target object label, $(o_0, ..., o_n)$ denote the names of the objects in the last image, BOS and EOS are begin of string and end of string tokens respectively. Consequently, $(i_1, ..., i_m)$ denotes the tokens that correspond to the instruction.

3.2. Llama2 Discriminator

Another Llama2 model that acts as a discriminator evaluates whether the generated instruction matches the visual trajectory and aligns with real human instructions. This component ensures that the generated instructions are realistic and contextually accurate. The purpose of the discriminator is to perform a classification task between real and fake instructions. Here, the ground truth instructions are referred to as real instructions, whereas the instructions generated by the Llama2 decoder are fake. Binary cross-entropy loss is used to minimize the error between the actual output and the generated output (real or fake).

3.3. Adversarial Training using Gumbel Softmax

The generator and discriminator are trained simultaneously in a competitive setup. The generator aims to produce instructions that are indistinguishable from ground-truth human instructions, fooling the discriminator. It minimizes a loss function on the basis of how "realistic" its outputs are judged to be. The discriminator is trained to differentiate between real human-written instructions and synthetic instructions generated by the model. It minimizes a binary cross-entropy loss that measures its ability to correctly classify instructions as real or fake. Gumbel-Softmax is used to make the discrete token generation process differentiable, enabling backpropagation through the generator during adversarial training.

The generator loss is defined as:

$$\mathcal{L}_G = -\log(D(I_G, \boldsymbol{x})), \tag{2}$$

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 loss is:

$$\mathcal{L}_D = -\log(1 - D(I_G, \boldsymbol{x})) - \log(D(I_R, \boldsymbol{x})), \tag{3}$$

where $I_R \in R(\boldsymbol{x})$ is the ground-truth instruction.

Table 1
Image description experiments comparison of AIGeN-Llama with AIGeN [3]

	Val Seen					Val Unseen				
Model	BLEU-1	METEOR	ROUGE	CIDEr	SPICE	BLEU-1	METEOR	ROUGE	CIDEr	SPICE
AIGeN	**48.4**	**22.8**	46.5	89.0	32.9	**42.1**	**17.9**	39.3	48.6	22.8
AIGeN-Llama	35.6	21.8	**53.8**	**117.6**	**41.3**	26.3	17.1	**44.4**	**81.9**	**33.2**

4. Experiments

We evaluated AIGeN-Llama on a widely used VLN dataset, REVERIE. In REVERIE, navigation sequences are composed of 360° images that are collected at the nodes of navigation graphs in Matterport3D environments [16]. Each navigation sequence requires agents to identify and interact with specific objects at the target location, adding complexity to the task. Only the frontal view of the 360° images, with a field of view of 60° is considered. For evaluation, we follow the standard split of training, validation seen, and validation unseen environments provided by the datasets. The training of AIGeN-Llama uses a learning rate of $0.2e-3$ for the generator and $0.2e-2$ for the discriminator, a batch size of 1, and Adam [17] as the optimizer. We use a pretrained Llama2 7B chat model for the generator and a pretrained Open Llama 3B model for the discriminator. The visual features used by the model are extracted using ResNet-152. Both the generator and the discriminator are individually trained before training them in an adversarial manner. This is done to ensure that the generator is already able to generate somewhat relevant instructions when trained together with the discriminator in an adversarial manner. Although the batch size is 1, we accumulate the gradients and update the optimizer every 48 steps. During the evaluation, the discriminator of the model is dropped, and the instructions are generated using the trained generator only.

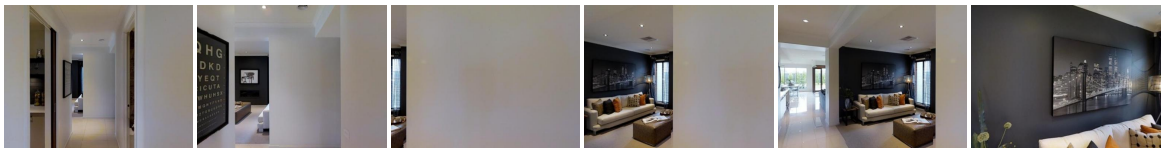
4.1. Quantitative Results

To evaluate the improvements introduced by AIGeN-Llama over its predecessor, AIGeN [3], we conduct a detailed comparison of the quality of generated instructions in terms of both descriptive richness and alignment with the input trajectories. The comparison focuses on two key aspects: instruction realism and contextual relevance to visual data. The comparison uses the standard image description metrics [18], namely BLEU [19], METEOR [20], ROUGE [21], CIDEr [22], and SPICE [23]. 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 [22]. As can be seen in Table 1, the metrics related to ROUGE, CIDEr, and SPICE are considerably higher for AIGeN-Llama than for AIGeN. Although AIGeN-Llama has lower BLEU and ROUGE scores compared to AIGeN, it's important to note that these metrics were originally designed for machine translation, where nearly exact word-for-word matches are expected. Low BLEU and METEOR scores alongside high CIDEr, ROUGE, and SPICE scores suggest that while the generated captions may not match the reference texts in wording or exact phrasing, they are capturing the core semantic content effectively.

4.2. Qualitative Results

Fig. 3 shows three qualitative samples in which the instructions generated by AIGeN-Llama are compared with the ground-truth instructions. All three samples have been taken from the "unseen" validation split of REVERIE, so that AIGeN-Llama has never seen these environments during training. The first two examples (a) and (b) are positive, while the latter is negative. In the first and second examples, both the goal rooms (dining room and living room) and the target objects (plant in both cases) are recognized

(a) **GT**: *Go to the dining room on level 1 with round table and center the plant on the table.*
AIGeN-Llama: *Go to the dining room and water the plant.*



(b) **GT**: *Enter the living room and pick up the potted plant.*
AIGeN-Llama: *Go to the living room and water the plant.*

(c) **GT**: *Pull out the second stool from the left side in the kitchen.*
AIGeN-Llama: *Go to the dining room and pull out the chair on your left.*

Figure 3: Sample image sequences from REVERIE Val Unseen split with corresponding ground-truth instruction and synthetic instructions generated using AIGeN-Llama. The images in each sequence have been reduced to 6 to facilitate the graphical presentation and we only show the frontal image of the panoramic observation at each timestep.

correctly. In the third example, 'kitchen' is recognized as a 'dining room' and 'stool' is recognized as a 'chair'. Looking at the last image of the trajectory (c), it is understandable that there is no clear boundary segregating the kitchen and the dining table. Moreover, 'chair' and 'stool' are quite close to each other in terminology, and hence, it is easy to confuse the two.

5. Conclusions and Future Works

In this work, we introduced AIGeN-Llama, a novel adversarial framework for generating high-quality, and realistic instructions in VLN. Using the advanced generative and discriminative capabilities of the Llama2 language model, AIGeN-Llama addresses key limitations of previous works, including excessive reliance on human-annotated data. The adversarial setup, where Llama2 serves as both a generator and a discriminator, enables the generation of synthetic instructions that closely align with human-authored text while maintaining descriptive precision. Our experiments demonstrate that AIGeN-Llama outperforms previous models like AIGeN on multiple evaluation metrics, namely ROUGE, CIDEr, and SPICE. This shows that AIGeN-Llama is capable of capturing the core semantic content effectively. In the future, we would like to test if the AIGeN-Llama helps to improve the navigation performance.

Acknowledgments

The authors were supported by Marie Sklodowska-Curie Action Horizon 2020 (Grant agreement No. 955778) for the project "Personalized Robotics as Service Oriented Applications" ("PERSEO") and "Fit for Medical Robotics" ("Fit4MedRob") project, funded by the Italian Ministry of University and Research.

References

[1] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, I. Laptev, Learning from Unlabeled 3D Environments for Vision-and-Language Navigation, in: Proceedings of the European Conference on Computer Vision, 2022.

[2] P.-L. Guhur, M. Tapaswi, S. Chen, I. Laptev, C. Schmid, Airbert: In-Domain Pretraining for Vision-and-Language Navigation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[3] N. Rawal, R. Bigazzi, L. Baraldi, R. Cucchiara, Aigen: An adversarial approach for instruction generation in vln, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 2070–2080.

[4] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, A. 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.

[5] X. Lin, G. Li, Y. Yu, Scene-Intuitive Agent for Remote Embodied Visual Grounding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[6] F. Landi, L. Baraldi, M. Cornia, M. Corsini, R. Cucchiara, Multimodal Attention Networks for Low-Level Vision-and-Language Navigation, Computer Vision and Image Understanding (2021).

[7] S. Chen, P.-L. Guhur, C. Schmid, I. Laptev, History Aware Multimodal Transformer for Vision-and-Language Navigation, Advances in Neural Information Processing Systems (2021).

[8] Y. Hong, C. Rodriguez, Y. Qi, Q. Wu, S. Gould, Language and Visual Entity Relationship Graph for Agent Navigation, Advances in Neural Information Processing Systems (2020).

[9] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, I. 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.

[10] D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L.-P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, T. Darrell, Speaker-Follower Models for Vision-and-Language Navigation, Advances in Neural Information Processing Systems (2018).

[11] A. Kamath, P. Anderson, S. Wang, J. Y. Koh, A. Ku, A. Waters, Y. Yang, J. Baldridge, Z. 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.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, Generative Adversarial Nets, in: Advances in Neural Information Processing Systems, 2014.

[13] B. Dai, S. Fidler, R. Urtasun, D. Lin, Towards Diverse and Natural Image Descriptions via a Conditional GAN, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017.

[14] R. Shetty, M. Rohrbach, L. Anne Hendricks, M. Fritz, B. 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.

[15] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, R. Girdhar, Masked-attention Mask Transformer for Universal Image Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[16] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, Y. Zhang, Matterport3D: Learning from RGB-D Data in Indoor Environments, in: Proceedings of the International Conference on 3D Vision, 2017.

[17] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, Proceedings of the International Conference on Learning Representations (2014).

[18] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, R. Cucchiara, From Show to Tell: A Survey on Deep Learning-based Image Captioning, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

[19] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2002.
[20] S. Banerjee, A. 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.
[21] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshops, 2004.
[22] R. Vedantam, C. Lawrence Zitnick, D. Parikh, CIDEr: Consensus-based Image Description Evaluation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015.
[23] P. Anderson, B. Fernando, M. Johnson, S. Gould, SPICE: Semantic Propositional Image Caption Evaluation, in: Proceedings of the European Conference on Computer Vision, 2016.