CROSS-LINGUAL VISION-LANGUAGE NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Navigation (VLN) is the task where an agent is commanded to navigate in photo-realistic unknown environments with natural language instructions. Previous research on VLN is primarily conducted on the Room-to-Room (R2R) dataset with only English instructions. The ultimate goal of VLN, however, is to serve people speaking arbitrary languages. To do this, we collect a cross-lingual R2R dataset, which extends the original benchmark with corresponding Chinese instructions. But it is time-consuming and expensive to collect large-scale human instructions for every existing language. Based on the newly introduced dataset, we propose a general cross-lingual VLN framework to enable instruction-following navigation for different languages. We first explore the possibility of building a cross-lingual agent when no training data of the target language is available. The cross-lingual agent is equipped with a meta-learner to aggregate cross-lingual representations and a visually grounded cross-lingual alignment module to align textual representations of different languages. Under the zero-shot learning scenario, our model shows competitive results even compared to a model trained with all target language instructions. In addition, we introduce an adversarial domain adaption loss to improve the transferring ability of our model when given a certain amount of target language data. Our methods and dataset demonstrate the potentials of building a cross-lingual agent to serve speakers with different languages.

1 INTRODUCTION

Recently, the Vision-Language Navigation (VLN) task (Anderson et al., 2018), which requires the agent to follow natural language instructions and navigate in houses, has drawn significant attention. In contrast to some existing navigation tasks (Mirowski et al., 2016; Zhu et al., 2017), where the agent has an explicit representation of the target to know if the goal has been reached or not, an agent in the VLN task can only infer the target from natural language instructions. Therefore, in addition to normal visual challenges in navigation tasks, language understanding and cross-modal alignment are essential to complete the VLN task.

However, existing benchmarks (Anderson et al., 2018; Chen et al., 2019) for the VLN task are all monolingual in that they only contain English instructions. Specifically, the navigation agents are trained and tested with only English corpus and thus unable to serve non-English speakers. To fill this gap, one can collect the corresponding instructions in the language that the agent is expected to execute. But it is not scalable and practical as there are thousands of languages on this planet and collecting large-scale data for each language would be very expensive and time-consuming.

Therefore, in this paper, we study the task of cross-lingual VLN to endow an agent the ability to understand multiple languages. First, *can we learn a model that has been trained on existing English instructions but is still able to perform reasonably well on a different language (e.g. Chinese)?* This is indeed a zero-shot learning scenario where no training data of target language is available. An intuitive approach is to train the agent with English data, and at test time, use a machine translation system to translate the target language instructions to English, which are then fed into the agent for testing (see the above part of Figure 1). The inverse solution is also rational: we can translate all English instructions into the target language and train the agent on the translated data, so it can be directly tested with target language instructions (see the below part of Figure 1). The former agent is tested on translated instructions while the latter is trained on translated instructions. Both solutions suffer from translation errors and deviation from the corresponding human-annotated instructions.

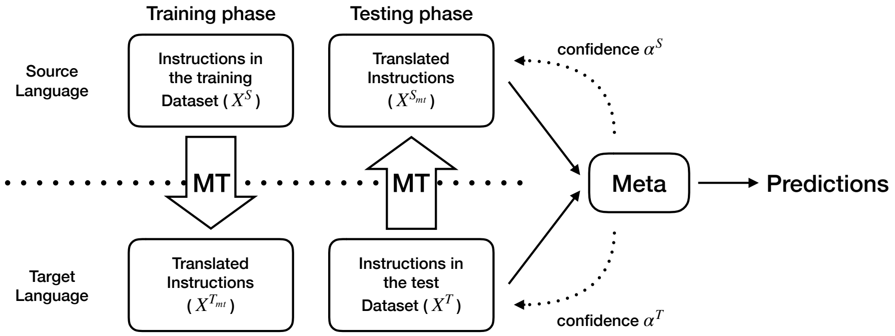

Training phase
Testing phase
Source Language
Instructions in the training Dataset (X^S)
Translated Instructions ($X^{S_{mt}}$)
confidence α^S
MT
MT
Meta
Predictions
Target Language
Translated Instructions ($X^{T_{mt}}$)
Instructions in the test Dataset (X^T)
confidence α^T

Figure 1: Overview of Meta Learner that learns to benefit from two learning schemes.

But meanwhile, the former is trained on human-annotated English instructions (which we view as "golden" data) and the latter is tested on "golden" target language instructions. Motivated by this fact, we design a cross-lingual VLN framework that learns to benefit from both solutions. As shown in Figure 1, we combine these two principles and introduce a **meta-learner**, which learns to produce beliefs for human-annotated instruction and its translation pair and dynamically fuse the cross-lingual representations for better navigation.

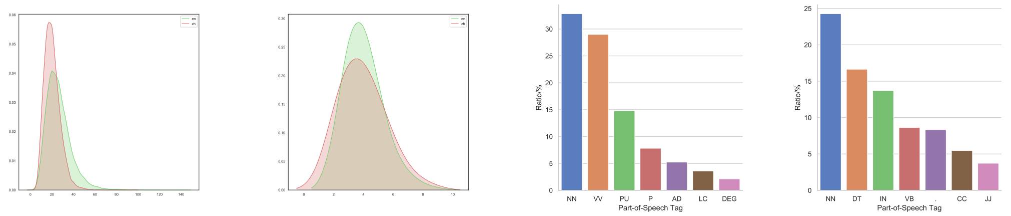
In this case, however, the training and inference are mismatched. During training, the agent takes source human language and target machine translation (MT) data as input, while during inference, it needs to navigate with target human instructions and source MT data. To better align the source and target languages, we propose a **visually grounded cross-lingual alignment** module to align the paired instructions via the same visual feature because they are describing the same demonstration path. The cross-lingual alignment loss can also implicitly alleviate the translation errors by aligning the human language and its MT pair in the latent visual space.

After obtaining an efficient zero-shot agent, we investigate the question that, *given a certain amount of data for the target language, can we learn a better adaptation model to improve source-to-target knowledge transfer?* The meta-learner and visually grounded cross-lingual alignment module provide a foundation for solving the circumstances that the agent has access to the source language and (partial) target language instructions for training. To further leverage the fact that the agent has access to the target language training data, we introduce an **adversarial domain adaption** loss to alleviate the domain shifting issue between human-annotated and MT data, thus enhancing the model's transferring ability.

To validate our methods, we collect a cross-lingual VLN dataset (XL-R2R) by extending complimentary Chinese instructions for the English instructions in the R2R dataset. Overall, our contributions are four-fold: (1) We collect the first cross-lingual VLN dataset to facilitate navigation models towards accomplishing instructions of various languages such as English and Chinese, and conduct analysis between English and Chinese corpus. (2) We introduce the task of cross-lingual vision-language navigation and propose a principled meta-learning method that dynamically utilizes the augmented MT data for zero-shot cross-lingual VLN. (3) We propose a visually grounded cross-lingual alignment module for better cross-lingual knowledge transfer. (4) We investigate how to transfer knowledge between human-annotated and MT data and introduce an adversarial domain adaption loss to improve the navigation performance given a certain amount of human-annotated target language data.

2 Problem Formulation

The cross-lingual vision-language navigation task is defined as follows: we consider an embodied agent that learns to follow natural language instructions and navigate from a starting pose to a goal location in photo-realistic 3D indoor environments. Formally, given an environment $\mathcal{E}$, an initial pose $p_1 = (v_1, \phi_1, \theta_1)$ (spatial position, heading, elevation angles) and natural language instructions $x_{1:N}$, the agent takes a sequence of actions $a_{1:T}$ to finally reach the goal G. Thus the VLN dataset $\mathcal{D}'$ is defined as $\{(\mathcal{E}, p_1, x_{1:N}, G)\}^{|\mathcal{D}'|}$. Note that we eliminate the footscript here for simplicity. At each time step t, the agent at pose p_t receives a new observation $\mathcal{I}_t = \mathcal{E}(p_t)$, which is a raw



(a) (b) (c) (d)

Figure 2: XL-R2R dataset analysis. (a) is instruction length distribution. (b) is sub-instruction number per instruction distribution. (c) and (d) are top 7 part-of-speech tag distribution of Chinese and English instructions.

RGB image pictured by the mounted camera. Then it takes an action a_t and leads to a new pose $p_{t+1} = (v_{t+1}, \phi_{t+1}, \theta_{t+1})$. Taking actions sequentially, the agent stops when a *stop* action is taken.

A cross-lingual VLN agent learns to understand multiple languages and navigate to the goal. Without loss of generality, we consider a bilingual situation coined as cross-lingual VLN. For this specific task, we built the XL-VLN dataset $\mathcal{D}$, which extends the VLN dataset and includes a bilingual version of instructions. Specifically, $\mathcal{D} = \{(\mathcal{E}, p_1, x^{\mathcal{S}}_{1:N}, x^{\mathcal{T}}_{1:N'}, G)\}^{|\mathcal{D}|}$, where $\mathcal{S}$ and $\mathcal{T}$ indicate source and target language domains separately. The source language domain $\mathcal{S}$ contains instructions in the source language covering the full VLN dataset $\mathcal{D}'$ (including training and testing splits), while the target language domain $\mathcal{T}$ consists of a fully annotated testing set and a training set in the target language that covers a varying percentage ϵ of trajectories of the training set in $\mathcal{D}'$ (ϵ may range from 0% to 100%). The agent is allowed to leverage both source and target language training sets and expected to perform navigation given an instruction from either the source or target language testing sets.

In this study, we first focus on a more challenging setting where no human-annotated target language data are available for training ($\epsilon = 0\%$), i.e., with no training data for the target language but the only access to the source language training set, the agent is required to follow a target language instruction $x^{\mathcal{T}}_{1:N'}$ to navigate in houses. Then we investigate the agent's transferring ability by gradually increasing the percentage of human-annotated target language instructions for training ($\epsilon = 0\%, 10\%, ..., 100\%$).

3 XL-R2R Dataset

We build a Cross-Lingual Room-to-Room (XL-R2R) dataset[1], the first cross-lingual dataset for the vision-language navigation task. The XL-R2R dataset includes 4,675/340/783 trajectories for train/validation seen/validation unseen sets, preserving the same split as in the R2R dataset. The official testing set of R2R is unavailable because the testing trajectories are held for challenge use. Each trajectory is described with 3 English and 3 Chinese instructions independently annotated by different workers.

Data Collection. We keep the English instructions of the R2R dataset and collect Chinese instructions via a public Chinese crowdsourcing platform. The Chinese instructions are annotated by native speakers through an interactive 3D WebGL environment, following the R2R dataset guidance (Anderson et al., 2018). More details can be found in the Appendix.

Data Analysis. XL-R2R dataset includes 5,798 trajectories in total and 17,394 instructions for both languages. The bilingual instruction part here is compared with each other from four perspectives for a broad understanding, including vocabulary, instruction length, sub-instruction number per instruction and part-of-speech tags. Removing words with less than 5 frequency, we obtain an English vocabulary with 1,583 words and a Chinese one with 1,134 words. The Chinese instructions are relatively shorter than English ones and less likely to be long sentences (Figure 2a). The instructions usually consist of several sub-instructions separated by punctuation tokens, and the number of sub-instructions per instruction distributes similar across language (Figure 2b). Figure 2c and

[1] XL-R2R dataset will be released.

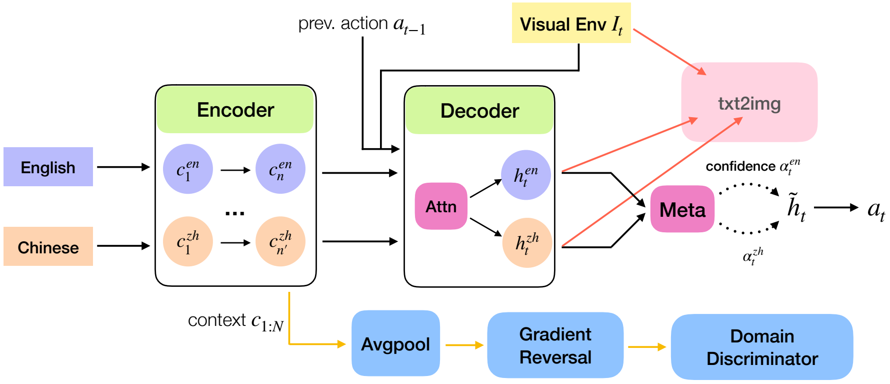


Figure 3: Architecture of the proposed cross-lingual VLN framework.

Figure 2d show that nouns and verbs, which often refer to landmarks and actions respectively, are more frequent in Chinese dataset (32.9% and 29.0%) than in English one (24.3% and 13.7%)[2].

4 Method

We present a general cross-lingual VLN framework in Figure 3. It is based on an encoder-decoder architecture and composed of three novel modules: a cross-lingual meta-learner (Meta), a visually grounded cross-lingual alignment module (txt2img), and an adversarial domain adaptation module. Particularly, as shown in Figure 3, both English and Chinese instructions are encoded by a shared encoder. Then the shared decoder takes the encoded contextual embeddings $c^{\mathcal{L}}_{1:N}$ from each language, the previous action a_{t-1}, and the local visual feature $\mathcal{I}_t$ as input, and produces hidden states h^{en}_t for English and h^{zh}_t for Chinese. The meta-learner learns to assign probabilities to h^{en}_t and h^{zh}_t, and then makes final predictions with the dynamically fused cross-lingual representation $\tilde{h}_t$.

In addition, the txt2img module is introduced to align h^{en}_t and h^{zh}_t in the visual space with the visual feature $\mathcal{I}_t$ as an anchor point, improving the cross-lingual knowledge transfer via a visually grounded cross-lingual alignment loss. The adversarial domain adaptation module is particularly designed for the transfer setting where human-annotated target language instructions are also provided. We employ a domain discriminator that is trying to distinguish human-annotated instructions from machine translation (MT) instructions, and a gradient reversal layer to reverse the gradients back-propagated to the encoder so that the encoder is indeed trying to generate indistinguishable representations for both human-annotated and MT data and align the distributions of the two domains.

4.1 Cross-lingual Meta-Learner

We employ the sequence-to-sequence architecture in Anderson et al. (2018) for both languages. Receiving a pair of natural language instruction $x^{\mathcal{L}}_{1:N}, \mathcal{L} \in \{\mathcal{S}, \mathcal{T}\}$, the agent encodes it with an embedding matrix followed by an LSTM encoder to obtain contextual word representations $c^{\mathcal{L}}_{1:N}$. The decoder LSTM takes the concatenation of current image feature $\mathcal{I}_t$ and previous action embedding a_{t-1} as input, and updates the hidden state s_{t-1} to s_t aware of the historical trajectory:

$$s_t = \text{LSTM}_{\text{dec}}(s_{t-1}, [\mathcal{I}_t, a_{t-1}]) \tag{1}$$

An attention mechanism is used to compute a weighted context representation, grounded on the instruction $c^{\mathcal{L}}_{1:N}$ by the hidden state s_t, then obtain final hidden states $h^{\mathcal{L}}_t$ for each language:

$$h^{\mathcal{L}}_t = \tanh(W[\tilde{c}^{\mathcal{L}}_t, s_t]) \tag{2}$$

$$\tilde{c}^{\mathcal{L}}_t = \text{Attn}(c^{\mathcal{L}}_{1:N}, s_t) \tag{3}$$

To bridge the gap between source and target languages, we leverage a machine translation (MT) system to translate the source language in the training data into the target language. During testing,

[2] The POS tags are obtained via Stanford Part-Of-Speech Tagger (Toutanova et al., 2003)

the MT system will translate the target language instruction into the source language. The MT data serves as augmented data for zero-shot or low-resource settings as well as associates two different human languages in general.

We take two instructions (the human language instruction and its MT pair) as input for both training and testing. We observed that, even if one instruction is a direct translation from the other, when the paired instructions are fed into the same encoder and decoder, the two instructions will often generate different predictions when executing. At each time step, when the agent observed the local visual environment, with two instructions at hand but lead to different next positions. It remains a challenging question which language representation the agent shall trust more.

Therefore, we propose a cross-lingual meta-learner that tries to help the agent make the judgment. At each time step, we let the cross-lingual meta-learner decide which language representation we should have more faith in, i.e., "learning to trust". The meta-learner is a SoftMax layer which takes the concatenation of two hidden states $h_t^{\mathcal{S}}$ and $h_t^{\mathcal{T}}$ as input, and produces a probability α_t representing the belief of the source language representation. The final hidden vector used for predicting actions is defined as a mixture of the representations in two languages:

$$\tilde{h}_t = \alpha_t h_t^{\mathcal{S}} + (1 - \alpha_t) h_t^{\mathcal{T}} \tag{4}$$

Finally, the predicted action distribution for the next time step is computed as:

$$P(a_t | a_{1:t-1}, \mathcal{I}_{1:t}, x_{1:N}^{\mathcal{S}}, x_{1:N}^{\mathcal{T}}) = \text{softmax}(\tilde{h}_t) \tag{5}$$

4.2 Aligning Cross-lingual Representations Through Visual Space

To better ground and align two languages to the images they describe, we map h_t^S and h_t^T into the latent space of image representations such that their similarity is maximized. In other words, we use the image space as an anchor point to align cross-lingual representations. Let $\mathcal{I}_t$ be the latent representation of the local visual environment on the target trajectory at time step t (e.g. the final layer of a ResNet), the loss function is formulated as:

$$L_{T2I} = \sum_{t \in 1:T} \left(\left\| \psi(W^{\mathcal{I}} \mathcal{I}_t) - \psi(W^{\mathcal{S}} h_t^{\mathcal{S}}) \right\|^2 + \left\| \psi(W^{\mathcal{I}} \mathcal{I}_t) - \psi(W^{\mathcal{T}} h_t^{\mathcal{T}}) \right\|^2 \right) \tag{6}$$

where ψ denotes a non-linearity activation such as ReLU or tanh, $W^{\mathcal{I}}$, $W^{\mathcal{S}}$ and $W^{\mathcal{T}}$ are the projection matrices. L2 distance is used to measure the similarity between contextual word and image features in the same vector space.

The intuition behind such an aligning mechanism is that, since the human instruction and the MT instruction are both describing the same trajectory, their representation should be close to the visual environment in some way (a projected latent space in our case). This module would ensure the consistency among cross-lingual representations and the visual inputs.

4.3 Transferring Knowledge via Adversarial Domain Adaptation

During training, we have human-annotated data for the source language and the machine-translated data for the target language, while an opposite situation for testing. To bridge the gap between training and inference, we leverage an adversarial domain adaption loss to make the context representations indistinguishable across domains for the transfer setting, where a certain amount of human-annotated instructions for the target language is available.

A sentence vector $\bar{c}$ is computed as the mean of the context vector $c_{1:N}$ to a single vector, then forwarded to a domain discriminator through the gradient reversal layer. With the gradient reversal layer, the gradients minimizing domain classification errors are passed back with opposed sign to the language encoder, which adversarially encourages the encoder to be domain-agnostic. We then minimize the following domain adaptation loss:

$$L_{domain} = -(y^{\mathcal{H}} \log \hat{y}^{\mathcal{H}} + (1 - y^{\mathcal{H}}) \log(1 - \hat{y}^{\mathcal{H}})) \tag{7}$$

where y^H is the domain label of an instruction indicating whether it is from human annotation or machine translation, and $\hat{y}^{\mathcal{H}}$ is the approximation by the domain discriminator.

	Validation Seen					Validation Unseen				
Model	PL	NE ↓	SR ↑	SPL ↑	CLS ↑	PL	NE ↓	SR ↑	SPL ↑	CLS ↑
train w/ MT	13.09	6.14	39.1	31.5	50.7	11.45	7.92	23.2	18.0	37.4
meta-learner	13.24	5.70	**44.4**	35.6	51.6	11.33	7.64	**25.4**	20.0	38.9
meta+txt2img	12.14	**5.66**	43.4	**36.4**	**53.7**	10.20	**7.60**	24.8	**20.5**	**40.5**
train w/ AN	12.69	5.52	45.3	37.3	53.2	10.89	7.71	26.2	21.5	39.7

Table 1: Performance comparison for zero-shot setting. Reported results are averages of 5 individual runs. *train w/ MT* indicates the model trained with Chinese MT data. *meta-learner* is our basic framework in Figure 1. *meta+txt2img* equips the meta-learner with txt2img module. The first three models are all for zero-shot learning. The last one, *train w/ AN* is to train the agent with 100% Chinese human annotated data. All models are tested with Chinese human instructions.

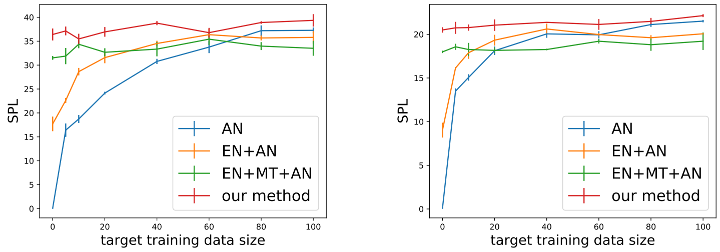
5 Experiments

Evaluation metrics. The following evaluation metrics are reported: (1) Path length (PL), which measures the total length of predicted paths. (2) Navigation Error (NE), mean of the shortest path distance in meters between the agents final location and the goal location. (3) Success Rate (SR), the percentage of final positions less than 3m away from the goal location. (4) Oracle Success Rate (OSR), the success rate if the agent can stop at the closest point to the goal along its trajectory. (5) Success rate weighted by (normalized inverse) Path Length (SPL), which trades-off Success Rate against trajectory length. (6) Coverage weighted by Length Score (CLS), recently introduced in (Jain et al., 2019), measures the fidelity to the described path, unlike previous metrics which are mostly based on goal completion.

5.1 Zero-shot Learning

We first report results for the zero-shot setting, to show the effectiveness of our two components, meta-learner, and txt2img. We compare with two models, a seq2seq model trained with human-annotated Chinese instructions (collected in XL-R2R dataset), and that trained with MT Chinese instructions translated from English. Results are shown in Table 1. **First**, there is a clear gap between training with human-annotated and MT data, indicating the insufficiency of using only an MT system for zero-shot learning. **Second**, our meta-learner can successfully aggregate the information of the annotated data and MT data, which enables efficient zero-shot learning. **Third**, the txt2img module further improves vision-language alignment, which especially helps the agent generalize better on the unseen data. Besides, even though the agent does not have access to the target annotation data, it achieves competitive results compared to training with 100% annotated data.

5.2 Transfer Learning



(a) Results on Validation Seen part (b) Results on Validation Unseen part

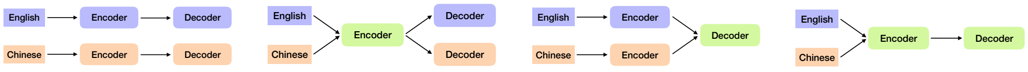
Figure 4: Learning curves of SPL on validation seen and unseen sets. *AN* is to train the model with Chinese annotations only. *EN+AN* and *EN+AN+MT* are to train the model with 100% English human data, or with 100% English human data plus Chinese MT data, with a certain amount of Chinese annotations to the training set. Our method is to enable knowledge transfer via adversarial learning, building on top of meta+txt2img.

To investigate the potential of transferring knowledge from English to Chinese, we draw learning curves by utilizing varying percentages of Chinese annotations for training (see Figure 4). The starting point is our zero-shot setting, where one has no access to the human-annotated data for the target language, and the endpoint is when one has 100% training data of the target language.

The figures demonstrate that the proposed adversarial domain adaption module provides consistent improvement over other methods, for both seen and unseen environments. The approach works for both low-resource and high-resource settings and is capable of transferring knowledge steadily as the size of target data grows. Besides, our transfer method trained with 40% Chinese human data can achieve similar performance as trained with 100% Chinese human data. This demonstrates the potential of building a functioning cross-lingual VLN agent by collecting a large-scale dataset for a certain language (i.e., English), and a small amount of data for other languages. One can also observe that pretraining with English data and MT Chinese data help the model learn useful encoding that is especially valuable when only limited Chinese training data are available.

5.3 Ablation Study on Parameter Sharing

To enable cross-lingual VLN, we examine four models (see Figure 5) equipped with our meta-learner and txt2img modules: (1) **Base-Bi**, which has two separate encoder-decoder. (2) **Shared Enc**, which has a shared language encoder. (3) **Shared Dec**, which has a shared policy decoder. (4) **Share Enc-Dec**, which shares both the encoder and the decoder, with different word embeddings for different languages. These models take English and Chinese natural language instructions as input, for both training and testing. They are also compared with **Base-mono**, which is a single encoder-decoder model trained and tested with Chinese human instructions only.



(a) Base-Bi (b) Shared Encoder (c) Shared Decoder (d) Shared Enc-Dec

Figure 5: Cross-lingual VLN models

		Validation Seen					Validation Unseen				
Model	#paras	PL	NE ↓	SR ↑	SPL ↑	CLS ↑	PL	NE ↓	SR ↑	SPL ↑	CLS ↑
Base-mono		12.69	5.52	45.3	37.3	53.2	10.89	7.71	26.2	21.5	39.7
Base-Bi	19.6M	12.74	5.67	43.6	35.9	51.8	10.684	7.408	27.7	22.5	41.0
Shared Enc	17.8M	13.12	5.64	44.5	36.4	52.0	11.050	7.395	27.7	22.5	41.0
Shared Dec	13.5M	12.13	5.06	51.0	43.7	**56.9**	11.069	**7.048**	31.0	**25.5**	**42.8**
Shared Enc-Dec	11.7M	12.62	**4.97**	**52.0**	**43.7**	56.7	11.187	7.080	**31.1**	24.9	42.1

Table 2: Performance comparison for cross-lingual VLN models. All models are trained and tested with English and Chinese annotation data. Results are averaged over 3 runs.

Table 2 shows the results of four architectures on the validation seen and unseen part. **First**, the performance of multi-lingual models are consistently improved over the monolingual model (*Base-mono*), indicating the potential of cross-lingual learning for improving navigation results. **Second**, sharing parameters can further boost navigation performance. **Finally**, *Shared Enc* and *Shared Enc-Dec* produce similar results, which motivates us to use a *Shared Enc-Dec* design since it yields a competitive good result with fewer parameters required.

5.4 Case Study

For a more intuitive understanding of the meta-learner, we visualize the confidences assigned to each language in Figure 6. In this case, the meta-learner trusts more on the human-annotated Chinese instruction which is of better quality. More specifically, at time step 10, when the meta-learner has the highest faith in the Chinese instruction, we visualize the textual attention on the whole instruction at this time step. Evidently, the corresponding textual attention on the Chinese command makes more sense than the machine-translated English command. The agent is supposed to keep turning left and then move forward to the green plant. The attentions on Chinese instruction assigns 0.25 to “turn left”, and nearly zero attention to “head towards the door” which is already completed by previous actions. While the attention on English is more uniform and less accurate than on Chinese.

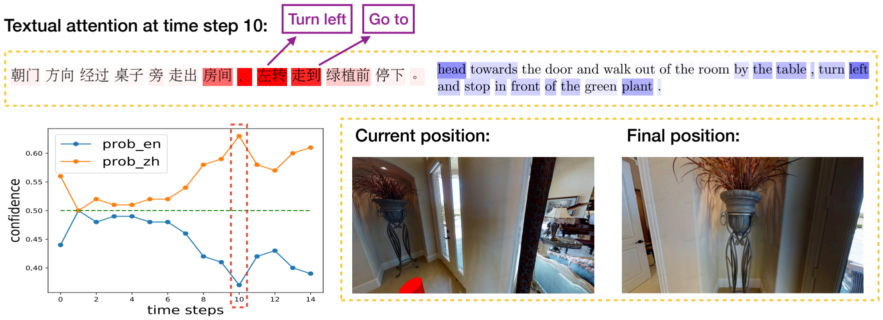


Figure 6: Case Study. We choose a succeeded instruction from the validation set for illustration.

6 Related Work

Vision and Language Grounding. Over the past years, deep learning approaches have boosted the performance of computer vision and natural language processing tasks (Krizhevsky et al., 2012; Sutskever et al., 2014; He et al., 2016; Vaswani et al., 2017). A large body of benchmarks are proposed to facilitate the research, including Image and Video caption (Lin et al., 2014; Krishna et al., 2017), VQA (Antol et al., 2015; Das et al., 2018), and visual dialog (Das et al., 2017). These tasks require grounding on both visual and textual modalities, but mostly limited to a fixed visual input. Thus, we focus on the task of vision-language navigation (VLN) (Anderson et al., 2018), where an agent needs to actively interact with the visual environment following language instructions.

Vision-Language Navigation. Several approaches have been proposed for the VLN task on the R2R dataset. For example, Wang et al. (2018) presented a planned-ahead module combining model-free and model-based reinforcement learning methods, Fried et al. (2018) introduced a speaker which can synthesize new instructions and implement pragmatic reasoning. Subsequent methods extend the speaker-follower model with Reinforced Cross-modal Matching (Wang et al., 2019a), self-monitoring (Ma et al., 2019), back-translation (Tan et al., 2019) etc. Previous works mainly improve navigation performance by data augmentation or leveraging efficient searching methods. In this paper, we address the task from a cross-lingual perspective, aiming at building an agent to execute instructions for different languages.

Cross-lingual Language Understanding. Learning cross-lingual representations is a crucial step to make natural language tasks scalable to all the world's languages. Recently, cross-lingual studies on typical NLP tasks has achieved success, such as Part-of-Speech tagging (Zhang et al., 2016; Kim et al., 2017), sentiment classification (Zhou et al., 2016) and Named Entity Recognition (Pan et al., 2017; Ni et al., 2017) These studies successfully disentangle the linguistic knowledge into language-common and language-specific parts and learn both knowledges with individual modules. Moreover, cross-lingual image and video captioning (Miyazaki & Shimizu, 2016; Wang et al., 2019b) aim to bridge vision and language towards a deeper understanding, by learning a cross-lingual model grounded on visual inputs. Our dataset and method address the cross-lingual representation learning for the vision-language navigation task. To our knowledge, we are the first to study the cross-lingual learning in a dynamic visual environment, where the agent needs to interact with its surroundings and take a sequence of actions.

7 Conclusion

In this paper, we introduce a new task, namely cross-lingual vision-language navigation, to study cross-lingual representation learning situated in the navigation task where cross-modal interaction with the real world is involved. We collect a cross-lingual R2R dataset and conduct pivot studies towards solving this challenging but practical task. The proposed cross-lingual VLN framework is proven effective in cross-lingual knowledge transfer. There are still lots of promising future directions for this task and dataset, e.g. to incorporate recent advances in VLN and greatly improve the model capacity. It would also be valuable to extend the cross-lingual setting to support numerous different languages in addition to English and Chinese.

REFERENCES

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 Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2018.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

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 Conference on Computer Vision and Pattern Recognition*, pp. 12538–12547, 2019.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 326–335, 2017.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2054–2063, 2018.

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. In *Advances in Neural Information Processing Systems*, pp. 3314–3325, 2018.

Yaroslav Ganin and Victor S Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv: Machine Learning*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Vihan Jain, Gabriel Magalhaes, Alex Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. *arXiv preprint arXiv:1905.12255*, 2019.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2832–2838, 2017.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-monitoring navigation agent via auxiliary progress estimation. *arXiv preprint arXiv:1901.03035*, 2019.

Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, et al. Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*, 2016.

Takashi Miyazaki and Nobuyuki Shimizu. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1780–1790, 2016.

Jian Ni, Georgiana Dinu, and Radu Florian. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. *arXiv preprint arXiv:1707.02483*, 2017.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 1946–1958, 2017.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. *arXiv preprint arXiv:1904.04195*, 2019.

Kristina Toutanova, D Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. pp. 173–180, 2003.

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*, pp. 5998–6008, 2017.

Xin Wang, Wenhan Xiong, Hongmin Wang, and William Yang Wang. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 37–53, 2018.

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 Conference on Computer Vision and Pattern Recognition*, pp. 6629–6638, 2019a.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. *arXiv preprint arXiv:1904.03493*, 2019b.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. Ten pairs to tag-multilingual pos tagging via coarse mapping between embeddings. Association for Computational Linguistics, 2016.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1403–1412, 2016.

Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3357–3364. IEEE, 2017.

A Implementation Details

We follow the same preprocessing procedure as in previous work. A ResNet-152 pretrained on ImageNet is used to extract image features, which are $2,048$-d vectors. Instructions are clipped with a maximum length of 80. Words are embedded into a 256-d vector space, and the action embedding is 32. The hidden size for the encoder and decoder LSTM is 512. The dropout ratio is 0.5. The meta-learner is a single fully connected layer. The dimension of vision-language alignment vector space is set to $1,024$. Each episode consists of no more that 40 actions.

The network is optimized via the ADAM optimizer with an initial learning rate of 0.001, a weight decay of 0.0005, and a batch size of 100. The learning rate of domain adaptation loss is scheduled with an adaption factor (Ganin & Lempitsky, 2014):

$$\lambda_p = \frac{2}{1 + \exp(-\gamma \cdot p)} - 1 \tag{8}$$

where γ is set to 10 and p is learning steps. We use $0.2\lambda_p$ to train the domain discriminator.

We run each model $30,000$ iterations and report the iteration with the highest SPL, and evaluate the models every 500 iterations.

B Results on English Test Set

For evaluating our proposed approach on the unseen test set, we participate in the Vision and Language Navigation challenge and submitted our results to the test server.

Here we treat Chinese as the source language and English as the target language. Hence for zero-shot learning, the agent has 100% Chinese annotated data but no English annotated data. The agent is commanded to follow English human instructions during testing. Results are shown in Table 3.

For zero-shot learning, our method (meta+txt2img) improves over the model trained with MT data only. For transfer learning, our method can efficiently transfer knowledge between Chinese and English data. The results are coherent with the reported results on the validation set. (See Table 1 and Figure 4).

	Test (unseen)					
Model	PL	NE ↓	OSR ↑	SR ↑	SPL ↑	Access to target training data
train w/ mt	12.40	8.12	0.325	0.224	0.172	✗
meta+txt2img	10.34	8.08	0.305	0.230	0.190	✗
train w/ an	10.92	7.47	0.344	0.259	0.212	✓
transfer	10.84	**7.50**	**0.348**	**0.265**	**0.222**	✓

Table 3: Performance comparison on the English test set. The first two rows are for zero-shot learning, the last two rows are trained with access to 100% target training data (i.e. English annotated instructions).

C Ablation Studies

Meta-learner To validate the effectiveness of the meta-learner, we compare it with a simple ensemble, which assigns equal confidence to two languages at all time steps, without any learnable parameters. The results are summarized in Table 4. Our meta-leaner has higher performance on the validation unseen set, suggests that "learning to trust" is important for cross-lingual vision-language navigation.

Adversarial Domain Adaptation Loss To demonstrate the domain adaptation loss indeed enhance knowledge transfer between two languages, we compare it with our vanilla zero-shot model, a meta-learner equipped with a txt2img module. Table 5 shows that, as the size of target training data grows, although the vanilla model can also benefit from the augmented data, the performance stops growing as the data size reaches 40% or 60%. Meanwhile, domain adaptation loss provides a more consistent

	Validation Seen					Validation Unseen				
Model	PL	NE ↓	SR ↑	SPL ↑	CLS ↑	PL	NE ↓	SR ↑	SPL ↑	CLS ↑
ensemble	12.542	5.71	44.1	**36.4**	**53.0**	10.90	7.72	24.2	19.4	38.7
meta-learner	13.24	**5.70**	**44.4**	35.6	51.6	11.33	**7.64**	**25.4**	**20.0**	**38.9**

Table 4: Ablation study for the meta-learner. Reported results are averages of 5 individual runs. *ensemble* is to assign equal weight to each language. *meta-learner* is our basic framework in Figure 1. Both results are reported on Chinese human instructions.

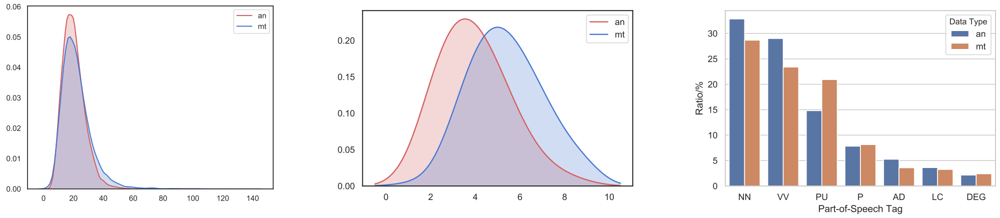
and steady improvement. At the endpoint (100%), the SPL is 22.14 vs 21.50, proves its efficiency and the potential of transferring knowledge between different languages.

Training data size (%)	5	10	20	40	60	80	100
meta+txt2img	20.52	20.71	20.56	21.29	**21.38**	21.00	21.50
meta+txt2img+domain	**20.74**	**20.77**	**21.05**	**21.37**	21.13	**21.48**	**22.14**

Table 5: Ablation study for domain adaptation. Reported results are averages of 3 individual runs on Chinese human instructions. *meta+txt2img* is our model for zero-shot setting. *meta+txt2img+domain* is to add the domain adaptation loss. The results are SPL values.

D Comparing Chinese Annotated Instructions with Machine Translated ones

We compare the statistics of the Chinese annotated dataset with a machine-translated one. The annotated instructions are more likely to contain fewer words as well as fewer instructions. Besides, nouns and verbs, which usually represent landmarks and actions in VLN task, are more frequent in annotated instructions than machine-translated ones.

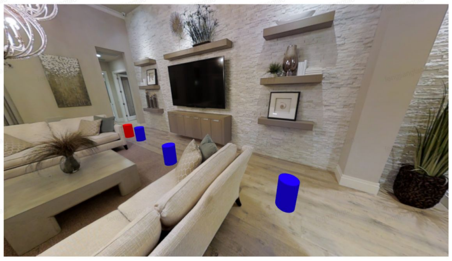


(a) (b) (c)

Figure 7: Statistics of human annotated and machine translated data. (a) is instruction length distribution. (b) is sub-instruction number per instruction distribution. (c) is top 7 part-of-speech tag distribution of annotated and machine translated instructions.

E CHINESE DATA COLLECTION INSTRUCTIONS

场景：



在标注中，你会看到如上图的场景。在场景中，你会看到许多圆柱，标记了你期望机器人前行的行进路线，其中红色圆柱表示最终的目的地，蓝色圆柱为轨迹的中间点，绿色表示起点（当前位置在起点故不可见）。请根据所给的路线用中文写下让机器人前进的路线的指令。

交互方式：

1. 左键单击并拖动可以观看全景，包括前后左右上下均可见；
2. 右键圆柱可以移动到圆柱所在的位置，并获得新的视野；

注意事项：

1. 机器人无法看到圆柱，请不要利用圆柱进行导航，圆柱仅为标注者提供路线参考；
2. 如果在图片中无法看到圆柱，请单击并拖拽观看全景图，寻找后续的轨迹；
3. 有时后目标在当前位置不可见时，例如在需要从一个房间到另一个房间时，可以通过指令移动到轨迹可见位置，如房间门口，再继续进行交互；
4. 你无法看见你所处位置的状态，即起点时无法见到绿色圆柱，终点则无法见到红色圆柱；
5. 在提供指令时，请使指令完整并反映出可供行走的轨迹的特征，切不要过于简单，如在上图中“到达门口”这类是不可取的；
6. 描述指令不必过分精确到圆柱，在其附近即可，只要大致展现出路径即可，即根据描述的指令能重演出前进路线即可；
7. 鼓励使用家具的位置进行导航，如“经过右边的窗户”之类的；
8. 请注意指令的语法，指令从简，描述精确，如“略微右转”、“一直向前走”