

Cross-Lingual Vision-Language Navigation

An Yan[♦], Xin Eric Wang[♦], Jiangtao Feng[♡], Lei Li[♡], William Yang Wang[♦]

[♦]UC San Diego, [♦]UC Santa Cruz, [♦]UC Santa Barbara

ayan@ucsd.edu, xwang366@ucsc.edu, william@cs.ucsb.edu

[♡]ByteDance AI Lab

{fengjiangtao, lileilab}@bytedance.com

Abstract

Commanding a robot to navigate with natural language instructions is a long-term goal for grounded language understanding and robotics. But the dominant language is English, according to previous studies on vision-language navigation (VLN). To go beyond English and serve people speaking different languages, we collect a cross-lingual Room-to-Room (XL-R2R) dataset, extending the original benchmark with new Chinese instructions. Based on this newly introduced dataset, we study how an agent can be trained on existing English instructions but navigate effectively with another language under a zero-shot learning scenario. Without any training data of the target language, our model shows competitive results even compared to a model with full access to the target language training data. Moreover, we investigate the transferring ability of our model when given a certain amount of target language training data.¹

1 Introduction

Grounded natural language understanding in the real world is an essential ability for a robot to communicate with humans (MacMahon et al., 2006; Chen and Mooney, 2011; Artzi and Zettlemoyer, 2013). The task of vision-language navigation (VLN) (Anderson et al., 2018b), which requires the agent to follow natural language instructions and navigate in houses, thrives recently due to photo-realistic simulation, free-form language instructions, and large-scale training. The VLN task is particularly challenging and requires an understanding of both language instructions and visual dynamics as well as cross-modal alignment.

Despite recent advances in VLN (Fried et al., 2018; Wang et al., 2019a; Tan et al., 2019), existing

VLN benchmarks (Anderson et al., 2018b; Chen et al., 2019) are all monolingual in that they only contain English instructions. 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 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 in the world, and collecting large-scale data for each language would be very expensive and time-consuming.

Therefore, in this paper, we introduce the task of cross-lingual VLN to endow an agent the ability to execute instructions in different languages. First, *can we learn an agent that is trained on existing English instructions but still able to navigate reasonably well for a different language?* This is essentially 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 upper 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 lower 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 human instructions. 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 cross-lingual lan-

¹XL-R2R is released at <https://github.com/zxslp/XL-VLN>

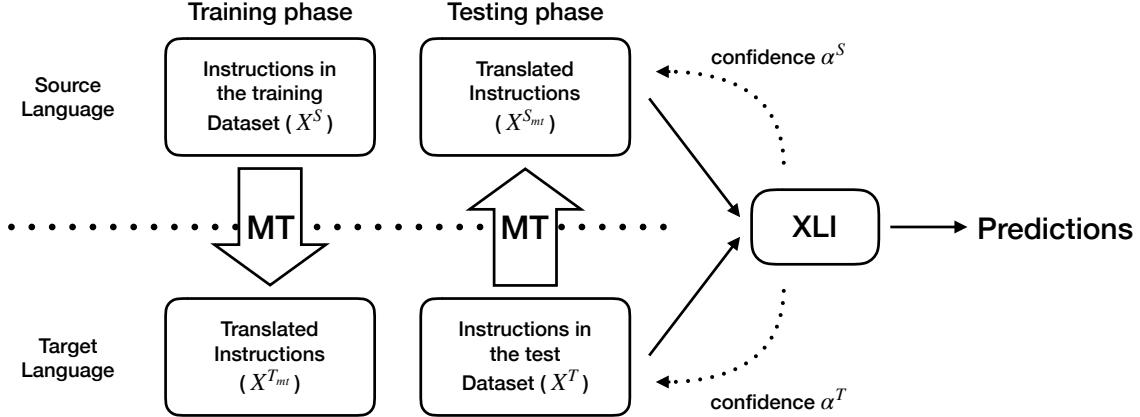


Figure 1: Overview of cross-lingual Instructor that learns to benefit from two learning schemes.

guage instructor (XLI), which learns to produce beliefs in the human instruction and its translation pair and to dynamically fuse the cross-lingual representations for better navigation.

After obtaining an efficient zero-shot agent, we investigate the question, *can our cross-lingual VLN framework improve source-to-target knowledge transfer if given a certain amount of data for the target language?* We conduct extensive experiments to show that the cross-lingual language instructor lay an effective foundation for solving the circumstances that the agent has access to the source language and (partial) target language instructions for training. To validate our methods, we introduce a cross-lingual VLN dataset by collecting complimentary Chinese instructions for the English instructions in the Room-to-Room dataset (Anderson et al., 2018b). Overall, our contributions are three-fold:

- We collect the first cross-lingual VLN dataset to facilitate navigation agents towards accomplishing instructions of various languages such as English and Chinese.
- We introduce the task of cross-lingual vision-language navigation and propose a principled cross-lingual learning framework with a pre-trained cross-lingual transformer and a cross-lingual language instructor.
- We demonstrate the efficiency of our model for cross-lingual knowledge transfer under two challenging settings, zero-shot learning where no target language data is available, and transfer learning where a certain amount of such data is given.

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

A cross-lingual VLN agent learns to understand different languages and navigate to the goal. Without loss of generality, we consider a bilingual situation coined as cross-lingual VLN. To support the task, we built the cross-lingual VLN dataset \mathcal{D} , which includes human instructions in two different languages. Specifically, $\mathcal{D} = \{(\mathcal{E}_i, p_{i,1}, x_{i,1:N}^S, x_{i,1:N}^T, G_i)\}_{i=1}^{|\mathcal{D}|}$, where S and T indicate source and target language domains. The footnote i is eliminated in future discussions for simplicity. Domain S contains instructions in the source language covering the full VLN dataset (including training and testing splits), while domain 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 vary from 0% to 100%). The agent is allowed to leverage both source and target language

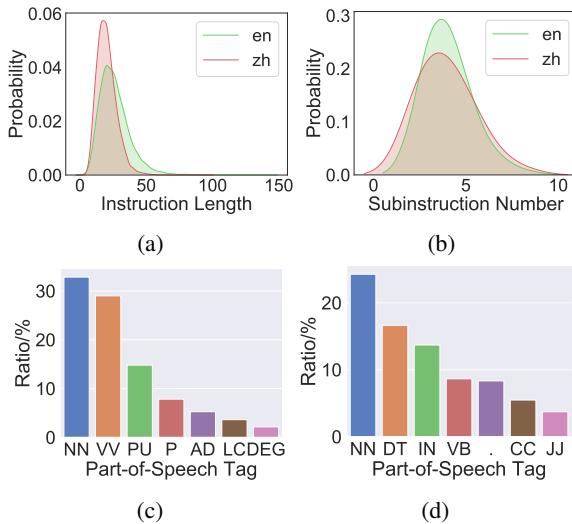


Figure 2: Analysis of XL-R2R English and Chinese corpora. (a) Distribution of instruction lengths. (b) Distribution of sub-instructions per instruction. (c) and (d) are distributions of part-of-speech tags for Chinese and English instructions.

training sets and expected to perform navigation given an instruction from either source or target language testing sets.

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

3 XL-R2R Dataset

We build a cross-lingual Room-to-Room (XL-R2R) dataset, the first cross-lingual dataset for the vision-language navigation task. It includes 4,675 trajectories for the *training* set, 340 for *validation seen*, and 783 for *validation unseen*, preserving the same split as the R2R dataset². Each trajectory is described with 3 English and 3 Chinese instructions independently annotated by different workers.

3.1 Data Collection

We keep the English instructions of the R2R dataset and collect Mandarin Chinese instructions via a public Chinese crowdsourcing platform. The Chi-

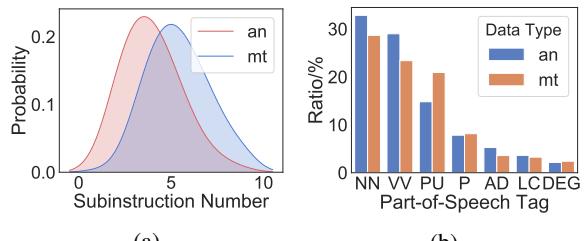


Figure 3: Statistics of human annotated and machine translated data. (a) is sub-instruction number per instruction distribution. (b) is top 7 part-of-speech tag distribution of annotated and machine translated instructions.

nese instructions are annotated by native speakers through an interactive 3D WebGL environment, following guidance by Anderson et al. (2018b). More details can be found in the Appendix.

3.2 Data Analysis

Chinese Annotations vs. English Annotations

XL-R2R includes 17,394 instructions in both English and Chinese, annotated on 5,798 trajectories in total. In Figure 2, we compare statistics of the English and Chinese corpora. Note that we segment Chinese words using the Jieba³ toolkit. Removing words with less than 5 frequency, we obtain an English vocabulary of 1,583 words and a Chinese vocabulary of 1,134 words. First, as shown in Figure 2a, Chinese instructions are shorter than English ones on average. Second, the instructions usually consist of several sub-instructions separated by punctuation, and we can observe that the numbers of sub-instructions per instruction distribute similar across languages (Figure 2b). Furthermore, Figure 2c and Figure 2d show that nouns and verbs, which often refer to landmarks and actions, are used more frequently in Chinese instructions (32.9% and 29.0%) than in English ones (24.3% and 13.7%)⁴.

Chinese Annotations vs. Machine Translations

In Figure 3, We compare the statistics of the Chinese annotated dataset with a machine-translated one. The annotated instructions are more likely to contain fewer sentences per instruction. Besides, nouns and verbs, which usually represent landmarks and actions in VLN task, are more frequent in annotated instructions than machine-translated

³<https://github.com/fxsjy/jieba>

⁴The POS tags are obtained via Stanford Part-Of-Speech Tagger (Toutanova et al., 2003).

²Note that the original testing set of R2R is unavailable because the testing trajectories are held for challenge use.

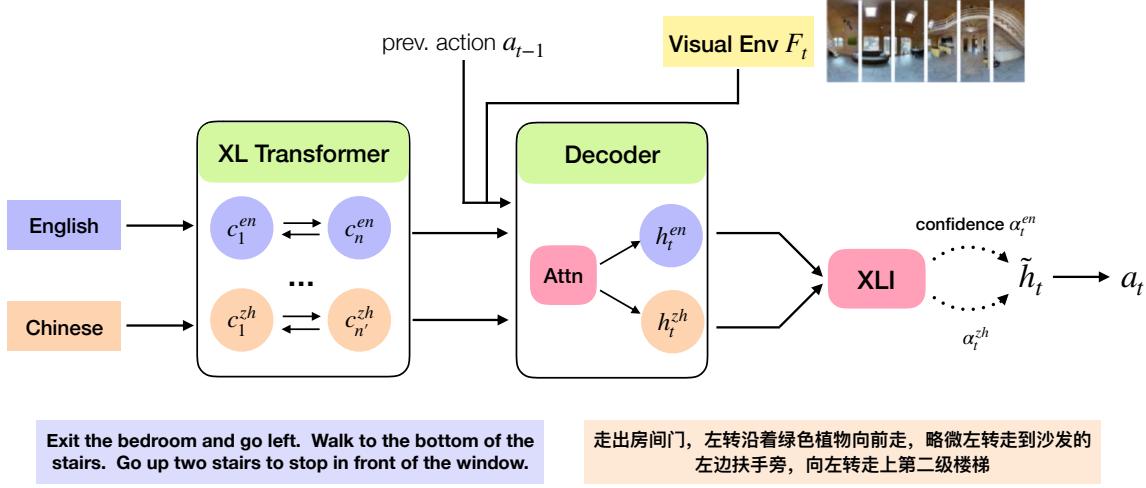


Figure 4: Illustration of the proposed cross-lingual VLN framework.

ones, which shares the same trend as comparing Chinese and English annotations. The above analysis shows that annotations and machine translations have different data distributions, hence directly deploying a model trained with machine translations will lead to performance decrease.

4 Method

We present a general cross-lingual VLN framework in Figure 4. It is composed of three modules: a pretrained cross-lingual transformer as the encoder, a decoder with panoramic action space and a cross-lingual language instructor (XLI). Particularly, as shown in Figure 4, both English and Chinese instructions are encoded by a cross-lingual transformer. Then the shared decoder takes the encoded contextual embeddings $c_{1:N}^{\mathcal{I}}$ from each language, the previous action a_{t-1} , and the local visual feature \mathcal{F}_t as input, and produces hidden states h_t^{en} for English and h_t^{zh} for Chinese. The language instructor learns to assign probabilities to h_t^{en} and h_t^{zh} , and then makes final predictions with the dynamically fused cross-lingual representation \tilde{h}_t .

Cross-Lingual Transformer Motivated by recent advances in pretraining multilingual language models at scale, we leverage a pretrained cross-lingual transformer to enable cross-lingual zero-shot and transfer learning. We employ the *XLM-R* architecture in Conneau et al. (2019) to encode both languages. It is a transformer model trained with multilingual masked language models (MLM) objective on 100 languages, which could address the data sparsity issue for our task with pretrained

multilingual knowledge.

Receiving a pair of natural language instruction $x_{1:N}^{\mathcal{I}}, \mathcal{I} \in \{\mathcal{S}, \mathcal{T}\}$, the cross-lingual transformer encodes the sentence to obtain contextual word representations $c_{1:N}^{\mathcal{I}}$, along with the pooled hidden state $h_{enc}^{\mathcal{I}}$ at the first token of the instruction.

Panoramic Decoder At each time step, the agent perceives a 360-degree panoramic view of its surrounding scene from its current location with image feature \mathcal{F}_t , discretized into 36 view angles. Each view angle is represented by an encoding vector v_i . The attended feature representation is computed with previous memory vector $s_t^{\mathcal{I}}$:

$$e_{t,i} = \text{Attn}_{vis}(\mathcal{F}_t, s_t^{\mathcal{I}}) \quad (1)$$

$$\mathcal{F}_{t,att} = \sum_i e_{t,i} \mathcal{F}_{t,i} \quad (2)$$

The decoder LSTM is initialized with $h_{enc}^{\mathcal{I}}$. It takes the concatenation of current attended image feature $\mathcal{F}_{t,att}$ and previous action embedding a_{t-1} as input, and updates the hidden state from $s_{t-1}^{\mathcal{I}}$ to $s_t^{\mathcal{I}}$ aware of the historical trajectory:

$$s_t^{\mathcal{I}} = \text{LSTM}_{dec}(s_{t-1}^{\mathcal{I}}, [\mathcal{F}_{t,att}, a_{t-1}]) \quad (3)$$

An attention mechanism is used to compute a weighted context representation, grounded on the instruction $c_{1:N}^{\mathcal{I}}$ by the hidden state $s_t^{\mathcal{I}}$, then obtain final hidden representation $h_t^{\mathcal{I}}$ for each language:

$$h_t^{\mathcal{I}} = \tanh(W[\tilde{c}_t^{\mathcal{I}}, s_t^{\mathcal{I}}]) \quad (4)$$

$$\tilde{c}_t^{\mathcal{I}} = \text{Attn}(c_{1:N}^{\mathcal{I}}, s_t^{\mathcal{I}}) \quad (5)$$

Cross-lingual Language Instructor To bridge the gap between source and target languages, we leverage a production-level machine translation (MT) system to translate the source language in the training data into the target language. During testing, 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. So we take two instructions (the human language instruction and its MT pair) as input for both training and testing. But we observed that these two instructions often generate different predictions, although one is the direct translation of the other. At each time step, when the agent observed the local visual environment, with two languages leading to different next positions, it remains a challenge which language representation to trust more.

Therefore, we propose a cross-lingual language instructor that learns to make the judgment. At each time step, we let the language instructor decide which language representation we should have more faith in, i.e., ‘‘learning to trust’’. The language instructor is a softmax layer which takes the concatenation of two hidden states h_t^S and h_t^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:

$$\tilde{h}_t = \alpha_t h_t^S + (1 - \alpha_t) h_t^T \quad (6)$$

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

$$P(a_t | a_{1:t-1}, \mathcal{F}_{1:t}, x_{1:N}^S, x_{1:N}^T) = \text{softmax}(\tilde{h}_t) \quad (7)$$

Thus the training objective is defined as the cross-entropy between the true actions and the predictive ones:

$$\mathcal{L}_{XLI} = - \sum_t \log P(a_t | a_{1:t-1}, \mathcal{F}_{1:t}, x_{1:N}^S, x_{1:N}^T) \quad (8)$$

5 Experiments

5.1 Experimental Setup

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 agent’s 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) (Anderson et al., 2018a), which trades-off Success Rate against trajectory length; (6) Coverage weighted by Length Score (CLS) (Jain et al., 2019), which measures the fidelity to the described path and is complementary to goal-oriented metrics.

Implementation Details. We follow the same preprocessing procedure as in previous work (Fried et al., 2018). A ResNet-152 model (He et al., 2016) pretrained on ImageNet is used to extract image features, which are 2048-d vectors. Instructions are clipped with a maximum length of 80. We use a XLM-RoBERTa base model (Conneau et al., 2019) pretrained on 100 languages as the encoder. The hidden sizes for encoder and decoder LSTM are 768 and 512 respectively. The dropout ratio is 0.5. Each episode consists of no more than 10 actions.

The network is optimized via the ADAM optimizer (Kingma and Ba, 2014) with initial learning rates of $1e-5$ on the pretrained encoder and $1e-4$ on the decoder, a weight decay of 0.0005, and a batch size of 100. We train each model for 10,000 iterations and evaluate it every 100 iterations. We report the iteration with the highest SPL on validation unseen set.

5.2 Zero-shot Learning

We report results under the zero-shot setting in Table 1, to show the effectiveness of our method. First, the difference between *train w/ MT* and *train w/ AN* validates our findings in Section 3.2, that mismatched distributions between machine translations and human annotations will result in performance decrease, which indicates the insufficiency of solely using MT data for zero-shot learning. Second, the clear gap between *train w/ MT* and *XLI* proves that our language instructor can successfully aggregate cross-representations of both human-annotated and MT data. Moreover, even though the agent does not have access to any annotated target language data, it achieves competitive results compared to *train w/ AN* that is trained with 100% annotated target language data.

Model	Validation Seen					Validation Unseen				
	NE ↓	OSR ↑	SR ↑	SPL ↑	CLS ↑	NE ↓	OSR ↑	SR ↑	SPL ↑	CLS ↑
train w/ MT	5.58	54.7	41.9 ± 6.0	35.8 ± 5.5	53.5 ± 3.7	6.99	39.2	28.9 ± 2.8	21.5 ± 2.7	36.8 ± 2.9
test w/ MT	5.53	57.1	43.5 ± 6.4	37.1 ± 6.3	54.3 ± 5.1	7.32	39.3	27.9 ± 4.7	20.4 ± 4.5	35.5 ± 4.1
XLI	5.12	59.0	48.4 ± 0.8	41.9 ± 1.2	57.7 ± 1.8	6.87	42.3	30.9 ± 0.5	23.4 ± 0.8	38.3 ± 1.0
train w/ AN	4.83	59.9	50.1 ± 0.6	43.2 ± 0.1	58.1 ± 0.3	6.91	41.1	31.4 ± 0.0	23.9 ± 0.2	38.4 ± 0.3

Table 1: Zero-shot learning results. Reported results are averages of 3 individual runs and shown with (*mean* \pm *std*). *train w/ MT* denotes the model trained with Chinese MT data. *test w/ MT* denotes the model trained with English annotations and test with English MT data translated from Chinese. *XLI* is the framework presented in Figure 1 that aggregates two learning schemes with a cross-lingual language instructor (Section 4). The first three models are all for zero-shot learning. The last one, *train w/ AN*, is trained with 100% human-annotated Chinese data. All models except *test w/ MT* are tested with human-annotated Chinese instructions.

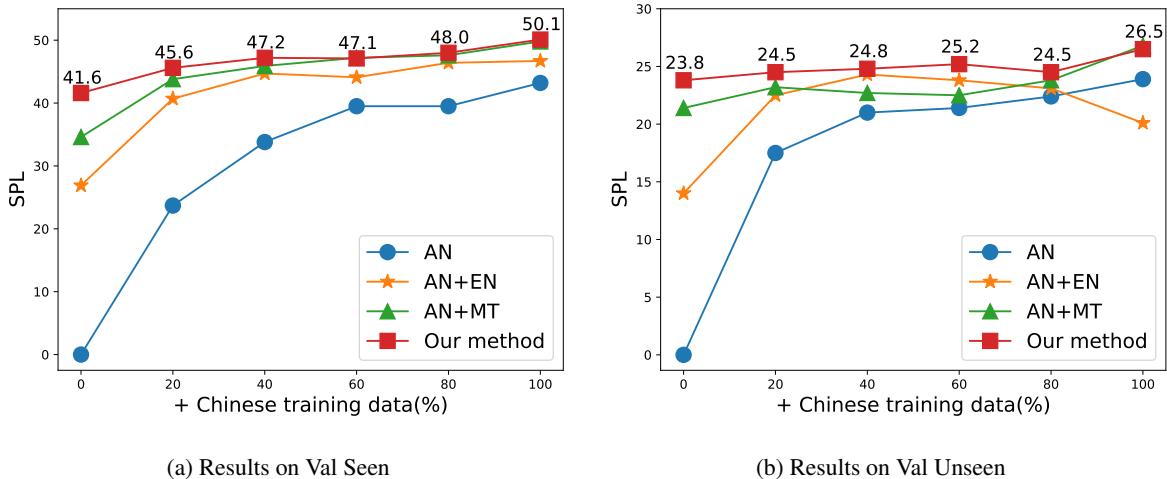


Figure 5: We examine the influence of using different percentages (from 0% to 100%) of target language (Chinese) instructions to train agents. We compare the results on XLI with three baselines. *AN*: only partial Chinese annotations (0% to 100%) are used for training. *AN+EN*: partial Chinese annotations and 100% English annotations. *AN+MT*: partial Chinese annotations and 100% Chinese translations of those English instructions.

5.3 Transfer Learning

To investigate the knowledge transfer effect from English to Chinese, we draw performance curves of utilizing varying percentages of Chinese annotations for training (see Figure 5). Particularly, the starting point is our zero-shot setting, where one has no access to human-annotated data of the target language (Chinese), and the endpoint is where one has 100% training data of the target language.

Figure 5 demonstrates that the proposed approach provides consistent improvements over other methods in both seen and unseen environments. First, our method works for both low-resource and high-resource settings, and improves the transferring ability steadily as the size of Chinese annotations grows. Besides, our method trained with 20% Chinese annotations has already outperformed the results as the model trained with 100% Chinese annotations. This demonstrates the

sample efficiency of our cross-lingual VLN model and the potential of scaling it for more languages with only a small amount of annotated data required. Finally, one can also observe that training with both English and MT Chinese data helps learn useful encoding that is especially valuable when only limited Chinese training data is available.

5.4 Encoder Variations

To enable cross-lingual VLN, we examine the navigation performance with different types of encoders as the backbone. As shown in Table 2, pretrained language models provide a better contextual representation of the instruction, hence leads to better navigation performance with the same decoder and training scheme. Furthermore, *M-BERT* and *XLM-R* have similar performance for our task. Vision-language navigation could serve as a new benchmark for evaluating these language models, i.e., in-

Encoder	Training data	Validation Seen					Validation Unseen				
		NE ↓	OSR ↑	SR ↑	SPL ↑	Δ SPL	NE ↓	OSR ↑	SR ↑	SPL ↑	Δ SPL
LSTM	MT	5.78	54.8	42.7	35.8	2.7	7.56	34.8	25.1	18.8	1.3
	AN	5.18	60.0	47.2	38.5		7.06	40.4	28.8	20.1	
M-BERT	MT	5.26	58.8	46.7	39.6	7.3	6.92	42.5	29.8	21.4	2.0
	AN	4.68	62.9	53.6	46.9		6.86	43.6	31.6	23.4	
XLM-R	MT	5.58	54.7	41.9	35.8	7.4	6.99	39.2	28.9	21.5	2.4
	AN	4.83	59.9	50.1	43.2		6.91	41.1	31.4	23.9	

Table 2: Performance comparison for different encoders. *LSTM* is to train the model with an LSTM encoder from scratch. *M-BERT* is an uncased multi-lingual BERT-base model pretrained with masked language modeling and Next sentence prediction on 102 languages (Devlin et al., 2018). *XLM-R* is a cross-lingual RoBERTa-base model, which is the encoder we mainly used for this task. ΔSPL is the difference of two SPLs trained with machine translations and human annotations.

Model	Test (unseen)				Access to target training data
	NE ↓	OSR ↑	SR ↑	SPL ↑	
train w/mt	7.3	37.0	28.3	21.4	✗
XLI (0%)	7.2	40.0	29.4	22.8	✗
train w/an	7.0	39.7	30.6	24.1	✓
XLI (100%)	6.7	42.8	33.0	25.2	✓

Table 3: Results on the R2R 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., annotated English instructions).

stead of common automatic metrics such as BLEU-4 (Papineni et al., 2002) or ROUGE-L (Lin, 2004), we can directly evaluate the performance of various language models with the navigation results. Finally, the gaps (ΔSPL) exist and are similar across all models when trained with different source data, i.e., machine translations and human annotations.

5.5 Results on English Test Set

We submitted the results to the VLN test server to evaluate the proposed approach on the unseen test set. We treat English as the target language for both zero-shot learning and transfer learning with 100% target training data. Results are presented in Table 3. For zero-shot learning, the agent has access to all human-annotated Chinese data but no English data during training. At test time, it is commanded to follow human-annotated English instructions. As shown in Table 3, our method (XLI) improves by 6.3% relatively over the model trained with MT data. For transfer learning, our method can efficiently transfer knowledge between Chinese and English data. The results here show similar trends as in the reported results on the Chinese validation set. (See Table 1 and Figure 5).

5.6 Case Study

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

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 (Chen and Mooney, 2011; 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; Xu et al., 2016), 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

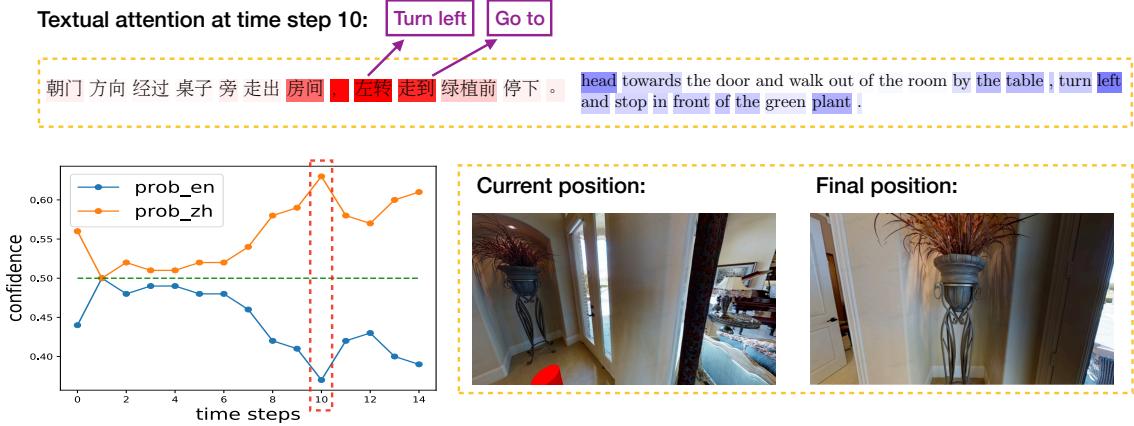


Figure 6: Case study. We choose a completed instruction from the validation set for illustration.

on the task of vision-language navigation (VLN) (Anderson et al., 2018b), 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. Cross-lingual studies on typical NLP tasks have achieved success, such as part-of-speech tagging (Zhang et al., 2016; Kim et al., 2017), sentiment classification (Zhou et al., 2016; Chen et al., 2018), named entity recognition (Pan et al., 2017; Ni et al., 2017) and vision-language tasks(Kim et al., 2019; Miyazaki and Shimizu, 2016; Wang et al., 2019b). These studies successfully disentangle the linguistic knowledge into language-common and

language-specific parts with individual modules.

Recently, bidirectional transformers (Devlin et al., 2018; Yang et al., 2019) pretrained on large-scale corpus data has drawn significant attention in the community. Its cross-lingual variants such as XLM and XLM-RoBERTa (Conneau and Lample, 2019; Conneau et al., 2019) showed superior performance on a wide range of down-stream cross-lingual transfer tasks. Our dataset and method address cross-lingual representation learning for the vision-language navigation task. To our knowledge, we are the first to study 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 and Future Work

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 shows its effectiveness 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 improve the model capacity. It would also be valuable to extend the dataset to support numerous different languages in addition to English and Chinese, as well as evaluate the cross-lingual performance of variant language models on this new benchmark.

References

- Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. 2018a. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*.
- Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1:49–62.
- David L Chen and Raymond J Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.
- Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.
- Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.
- Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2054–2063.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, pages 3314–3325.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. 2019. Stay on the path: Instruction fidelity in vision-and-language navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy.
- Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan A Plummer. 2019. Mule: Multimodal universal language embedding. *arXiv preprint arXiv:1909.03493*.
- Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan Al-Regib, Zsolt Kira, Richard Socher, and Caiming Xiong. 2019. Self-monitoring navigation agent via auxiliary progress estimation. *arXiv preprint arXiv:1901.03035*.
- Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI’06, page 1475–1482. AAAI Press.
- Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1780–1790.
- Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada.
- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. 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, pages 1946–1958.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.
- Hao Tan, Licheng Yu, and Mohit Bansal. 2019. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota.
- Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. 2019a. 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*, pages 6629–6638.
- Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019b. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *International Conference on Computer Vision (ICCV)*.
- Xin Wang, Wenhao Xiong, Hongmin Wang, and William Yang Wang. 2018. 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)*, pages 37–53.
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msrvtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.
- Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag-multilingual pos tagging via coarse mapping between embeddings. Association for Computational Linguistics.
- Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. 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)*, pages 1403–1412.

Appendix

A Chinese Data Collection

We paid 25 workers to do the data collection work via a public Chinese data collection platform, taking around 4 weeks to finish the task. The workers are paid reasonably, with an estimated hourly rate higher than the local minimum wage. Before starting annotation, we educated the workers in a face-to-face way with documented instructions to help them understand the task(see Figure 8).

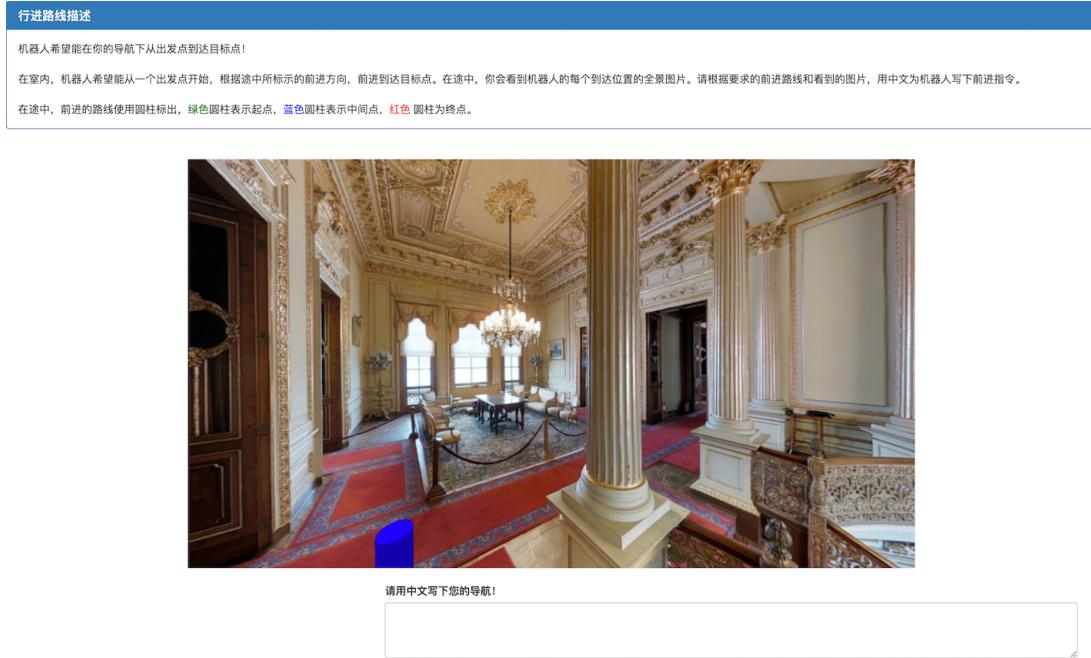


Figure 7: Interface for Chinese data collection.

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”

Figure 8: Instructions for Chinese data collection.