A Appendices

A.1 Analysis of Model Structure

In Fig. 4, we examine four model structures to evaluate the interactions between the two branches: (1) **Separate Enc-Dec** model, where two encoder-decoder models are trained separately for two branches. (2) **Shared Enc** model, which has a shared encoder but uses two different decoders for two branches. (3) **Shared Dec** model, which has different encoders for both linguistic and visual input but shared trajectory decoder. (4) **shared Enc-Dec** model, which shares both the encoder and the decoder. Note that this is the final architecture we use, which is demonstrated in Sec. 2. Table 4 shows the performance of four architectures on the development and test set. **First**, despite worse performance on other metrics, Separate Enc-Dec can achieve competitive performance on SPD and CLS against other two-branch shared models. The results show that the Separate Enc-Dec agent can produce high-fidelity trajectory matched with instruction but fail to stop at the correct location. This shows that to stop better, the stop indicator requires the information from the direction branch. **Second**, compared with Shared Enc model, Shared Dec performs competitively on SPD and CLS while much worse on other metrics, indicating that the stop branch learns better from the direction branch in the encoder phase. **Third**, both Shared Enc and Shared Dec show stronger ability to learn to stop; thus we use Shared Enc-Dec model, which requires fewer parameters. Improved performance shows the Shared Enc-Dec model learns to stop better than other architectures.

A.2 Hyper-Parameters Sensitivity Analysis

Threshold for Stop Signal We study the sensitivity of the threshold for stop signals on the development set. The result is shown in Fig. 5 (a). Task-Completion (TC) is consistent in a large range of thresholds, with a slight drop when the threshold is getting higher than 0.7 and sharp decreases when the threshold is close to 0 and 1. The results demonstrate that our approach is insensitive to the change of threshold for stop signals. The consistency of the performance means that the scores of stop signals are either low or high, rarely intermediate. This proves that our approach enables the agent to pay more attention to STOP; that is, the agent is cautious about deciding to stop and only stop when it is highly confident it reaches the goal.

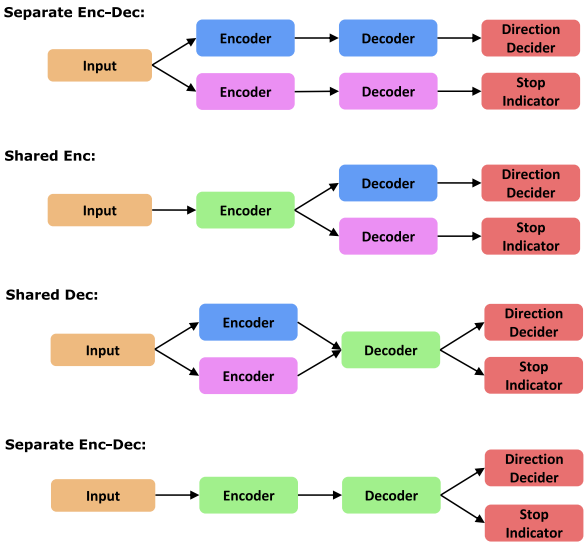


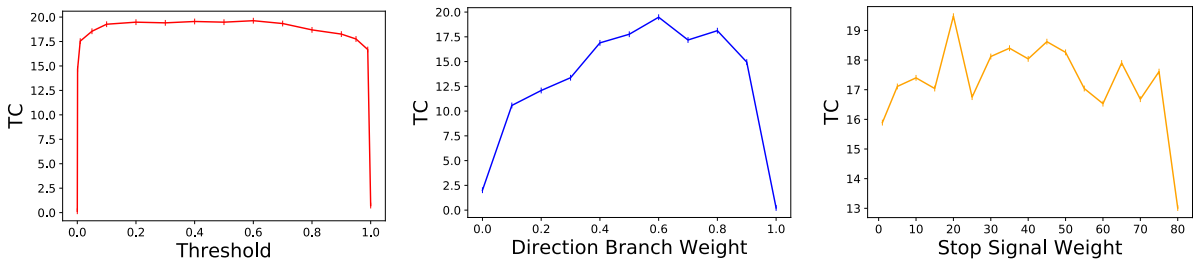
Figure 4: Two-branch VLN models. Input includes language instruction and local visual scene. one Encoder consists of a Visual Encoder and a Text Encoder in Fig. 2, and Decoder represents Trajectory Encoder in Fig. 2.

Direction Branch Weight We study the sensitivity of direction branch weight γ on the development set. The optimal value for γ is 0.6, as depicted in Fig. 5 (b), which demonstrates that the balance between the loss functions of two branch enables the agent to not only select correct directions at key points but also stop at the right place. As shown in the figure, smaller γ (0-0.5) results in relatively worse performance than higher γ, indicating that small γ enforces the agent to concentrate too much on STOP but ignore the choice for direction. Consistently good performance with larger γ (0.6-0.85) shows that only a small weight for the stop branch can significantly improve the agent's stop ability.

Stop Signal Weight We study the sensitivity of stop signal weight λ on the development set. As shown in the Fig. 5 (c), the optimal value for λ is 20. We can see that when $\lambda = 0$, our model's performance is similar to the ARC model (15.53 as shown in Table 1). However, when setting greater λ, the TC shows fluctuations, but is consistently better than ARC's performance. Only when λ increases to a large number of 80 does the performance decline sharply. This demonstrates the effectiveness of our proposed Weighted Cross-Entropy loss function, which consistently improves the agent's stop ability with a large range of λ.

Model	Development					Test				
	TC↑	SPD↓	SED↑	CLS↑	SDTW↑	TC↑	SPD↓	SED↑	CLS↑	SDTW↑
Separate Enc-Dec	13.71	17.67	13.35	55.24	13.32	14.14	**17.40**	13.71	54.56	13.61
Shared Dec	14.43	18.45	14.05	52.90	14.00	12.29	17.87	11.86	**54.86**	11.74
Shared Enc	18.75	18.19	18.32	52.42	18.27	15.55	18.31	15.21	52.87	15.19
Shared Enc-Dec	**19.48**	**17.05**	**19.02**	**55.68**	**18.97**	**16.68**	18.84	**16.34**	53.50	**16.34**

Table 4: Performance comparison for four different architectures of the two-branch model.



(a) (b) (c)

Figure 5: (a) Task Completion (TC) scores with different thresholds for the stop signal ($s_{t,2}$ in Equation 6). TC shows insensitivity to different thresholds. (b) TC scores with different direction branch weights γ in Equation 7. $\gamma = 0.6$ gives the highest TC. (c) TC scores with different stop signal weight λ in Equation 6. $\lambda = 20$ gives the highest TC. All the experiments are done on the development set,