We would like to thank all the reviewers for their insightful comments. All reviewers confirm the contribution of our paper in providing a novel framework for improving the generalisation of VLN tasks in addition to a sound formulation with wider range of applications. Please see below for specific responses.

Response to Reviewer #1:

Relation to Mixup: Mixup is not directly applicable to VLN since (1) it is sequential in nature, (2) an interpolation of state-action from one trajectory to another may lead to catastrophic difference in the objective. Our approach intervenes in the visual features to simulate the agent's behaviour in a counterfactual environment, where the agent still has to follow the same instruction and sequence of actions.

Difference with [34]: It is the closest approach to ours and general differences are highlighted in L89-93. Specifically for counterfactual distribution learning, similar to theirs, our approach creates counterfactual samples that are hard for the agent to follow with minimum interventions. However our approach is different to [34] in (1) ours is formulated to minimise the expected difference of the model on observations and the intervened samples while [34] resorted to importance sampling that could have a large variance (in fact it was not as successful in VLN); (2) ours is sequential; (3) our instructions have actions as opposed to vision and language only (4) we had a model of a speaker to incorporate.

Response to Reviewer #2:

Prior distribution and random noise: The hyper-parameters of the prior distribution have been selected through a simple grid search before applying the counterfactual distribution learning. Additionally, Environment Dropout [11] can be considered as a random noise injection which has been reported in Table 1 and 2.

Improvements: Generally, the reported improvements are indeed significant (around 4% in success rate and SPL), considering the fact that R2R task has been explored extensively in recent years so that even large-scale self-supervised pre-training [44] gains less than 4% improvement. By adding an insightful though simple counterfactual learning process, we gain further $1-2\%$ improvement on top of the improvement gained using the prior ($2-3\%$), which is a significant improvement on top of a model that enjoys a better generalisation than other baselines.

Counterfactual Learning: For the counterfactual distribution learning, we sample two pairs of real trajectories and only use language instruction of the first one. We will clarify this in the camera-ready version. In L167, we are approximating the marginalisation of $\mathbf{u}$ by adjusting the variable generated from the prior, instead of relying on costly methods like MCMC or a variational lower bound. It is also worth mentioning that as stated in L180 and according to Eq. 9, minimum edit (intervention) happens when $\mathbf{u}$ is close to one. In addition, for L194, please note that as stated in L178, $p(\mathbf{u} \mid \tau, \mathbf{c}) \propto p(\mathbf{u})\tilde{\pi}_{\boldsymbol{\theta}}(\tau \mid \mathbf{c}, \mathbf{u})$. Therefore, $p(\mathbf{u} \mid \tau, \mathbf{c})$ is maximised when $\mathbf{u}$ is close to one. Rather than integrating $\mathbf{u}$ out, we choose the most likely sample corresponding to a counterfactual. It should be noted that Figure 2.a in supplements is revealing the general SCM in VLN. Starting from a prior distribution, this variable is learnt conditioned on the inventions on the real observation and then is used for the counterfactual generation (Figure. 2.b).

State variable: In our experiments, $\mathbf{s}_t$ is the hidden state of the RNN model and the parameters are omitted for brevity.

Training setting: We freeze the speaker model during the counterfactual distribution learning. Additionally, as stated in L231, following the same setting as [11] and [8] for a fair comparison, we use teacher-forcing during IL.

Response to Reviewer #3:

Computational cost: Thanks for mentioning an interesting point. The computational cost of our proposed method highly depends on the number of iterations for learning the counterfactual distribution. Higher values result in better approximations of the exogenous variable while increasing the training time. We undertook several experiments for finding a point of best trade-off and found that even few iterations (5 as reported in Section 3.1 in the supplements) could contribute to a good understanding of the variable and therefore better results.

Sampling for RL: As indicated in Algorithm. 1, after finding the optimum value of $\mathbf{u}$ using $\{\tau, \mathbf{c}, \tau'\}$ and Eq. 11, at each step of the policy rollout, we intervene the observation using the learned $\mathbf{u}$ and τ'.

Response to Reviewer #4:

Computational cost: Please refer to the response to reviewer #3.

The effects of counterfactual data: We have demonstrated some trajectories revealing the difference between the baseline and our model over the same starting point and the same instruction in supplements. In addition, using our approach we observe a significant improvement in the test set that highlights the positive effect of our approach.

State variable: Please refer the response to reviewer #2. We will clarify in the camera-ready.

Equation 11: Thanks for pointing out the typo. We will replace a with a_t in the camera-ready version.

Algorithm 1: In fact, we use the optimised $\mathbf{u}$ to generate a new τ based on Eq. 9 for IL gain calculation. We haven't repeated the counterfactual generation for brevity of the algorithm.

Equation 4: As is mentioned in L182-185, for simplicity and efficiency, $\mathbf{u}$ is marginalised out from the Eq. 4 and is approximated using Eq. 11.