Clarify models and equations: **(R2)** Eq 1: The recurrent model contains an LSTM and a language graph. The global attended context $\tilde{h}^g_{t-1}$ produced by the language graph is fed to the LSTM as a language-aware "hidden state". The output of LSTM is fed back to the language graph as recurrent hidden state. **(R2)** Eq 1: We have experimented with feeding the object features to the LSTM but the agent performance degenerates. One possible reason is that Eq 1 keeps track of the navigation process and the completion of the instruction, but the same objects can appear in many viewpoints and disturb the tracking. **(R2)** Line 164: Empirically, we found that the temporal-link on the action node has the strongest influence on the agent's performance, compared with having the temporal-link on either the scene or the object node, or on more than one visual node. We believe this is because the temporal-link on the action node explicitly informs the visual graph about the action performed by the agent at the previous step, which could be used by the visual graph to infer a better action at the current step. **(R3)** Eq 10: Following previous work, we trained two models for R2R and R4R tasks respectively, using the same learning rate and optimizer as in the baseline method EnvDrop [28]. The coefficient λ is set to 0.2 in both tasks. **(R4)** Eq 2: The same h_t and u are fed into the attention modules with different learned parameters (see Appendix A.1 and [34]). The three attended contexts are used to initialize corresponding visual node features. Hence the three attended contexts will be relevant for three different visual clues.

Complexity (R1, R3): VLN is a complicated task, instead of having a monolithic network [9, 28], we add structure to our model to enhance its learning (better performance) and its interpretability. Introducing such structure/complexity in design actually reduces the complexity in training and inference: concurrent work PREVALENT [9] performs pre-training with 6,582K image-text-action triplets on eight V100 GPUs, but our work does not. Our model has about 22M parameters, which is larger than the baseline model EnvDrop (13M) [28], but much smaller than PREVALENT (190M, roughly LXMERT+EnvDrop). As for inference time, our method is 1.37 times faster on average than EnvDrop on the validation unseen split, because our agent is more efficient in solving the navigation tasks.

Performance: **(R1)** In contrast with concurrent work that applies auxiliary reasoning tasks (AuxRN [36]) or performs pretraining (PREVALENT [9]), our method without applying those training techniques still achieves better results in NE, SR and SPL metrics on the R2R test split. **(R3)** The ranking on leaderboard does not distinguish different experiment settings (e.g., single-run, beam-search, pre-exploration or multiple-instructions). Our method achieves the best SPL under the single-run setting (the primary setting for VLN)(see Line 236 and Table 1 caption). **(R4)** We follow the common practice of reporting the best performing model on the validation-unseen split during training [28,26,9] and evaluate it on the test split. We agree with the reviewer to analyze the failure modes, especially given our findings in the Ablation study. We will add some examples of failure cases to the Appendix of our paper. **(R4)** As in some recent work [9,19], we did not evaluate our agent with pre-exploration, which is not an original setting in VLN. However, we agree with R4 that learning a visual graph for the pre-explore environments and incorporate the pre-learned graph into training is a promising direction. We will consider it in our future work.

Contribution and Theory (R3): Based on our observation of the given instructions, in theory, if an agent can identify the three visual clues in the environment and learn about their relationships, it is more likely that it can successfully interpret the complex instructions and correctly perceive the environment. Therefore, we purpose a novel graph network for VLN to capture and utilize the inter- and intra-modal relationships among language and visual entities. We empirically validate our theory by evaluating our method on R2R and R4R benchmarks. We show the importance of the visual clues through our qualitative analysis, and the effectiveness of the two graph networks in the ablation studies.

Generalizability: **(R2)** We thank the reviewer's suggestion of evaluating our method on other VLN datasets such as Touchdown (street view) and CVDN (indoor dialog) to show the generalizability. Considering that our proposed relationship graph is a visual-textual encoder, it can be applied on these datasets with minor modifications. We will leave it as further work and add it to our (public) git repository. **(R4)** Our work is tailored for VLN because we want to solve VLN, which is a complicated and significant problem (and application). However, our work is not limited to VLN, e.g., it can be applied to temporal localization with languages in videos with some minor changes.

Ablation study and Future direction: **(R1, R4)** The removal of language graph means the message passing in the visual graph is not conditioned on the relational contexts. The removal of visual graph means there is no message passing (hence, the language graph has no effect). In both cases, the action prediction (Eq 9) is conditioned on all the existing visual node features. From model #4, #5 and Full, the object features give a large boost in performance when the two graph networks exist. **(R2)** Line 280: Our observation here was made during preliminary investigations and not part of our main result. However, we believe the finding is important and we would like to share it with other researchers, we will put the result on the final version of our paper or to our git repository. **(R4)** Self-supervised learning: As explained in Line 231-232, we applied the self-supervised learning as in EnvDrop [28], which uses a trained speaker [8,28] to generate instructions for randomly sampled trajectories for data augmentation.

We thank all the reviewers for acknowledging the contribution and strengths of our paper, as well as the thoughtful comments and the constructive suggestions. We will clarify all the unclear points and add the missing reference in the final version of our paper.