We heartily thank all the reviewers for their thoughtful comments and valuable feedback. We will try to address your queries and clarify any misunderstandings below.

To R1:

1. We have not used the word "motion module" in the paper, so it is slightly unclear which sub-module you are refering to. From the comments, our best guess is that you are referring to the linear velocity estimator (eq 2) and the intentional motion LSTM (eq 3) jointly as motion module. To highlight the improvement caused by the FQA interaction module, we show another ablation dropping the interaction module completely by setting all attention vectors to 0, while keeping the linear velocity estimate and the intentional motion LSTM (eqs 2, 3) intact. The resulting RMSEs shown below capture the severe drop in performance on all datasets, thereby showing that the improvement indeed comes from the interaction module.

ETH-UCY	Collisions	NGsim	Charges	NBA
0.549 ± 0.006	0.236 ± 0.0003	5.756 ± 0.152	0.523 ± 0.001	6.038 ± 0.044

2. Having binary decisions would make the architecture non-differentiable. Alternatively one maintains bernoulli decision RVs and infers their distributions using variational inference while making predictions. This is exactly what our NRI baseline does (and it is outperformed by FQA).

3. When removing the "decision making", we replace all sub-modules between the inputs of FQA module uptil V_{sr} (fig 2 in paper) and replace them with fully-connected layers with equivalent number of learnable params. A direct comparison with self-attention is not feasible since self-attention acts on a single entity, not a pair of entities (s and r). In fact, one way to interpret our FQA is as extending self-attention to pairs of entities.

To R2:

Thank you for the excellent suggestions! Indeed FQA can be easily extended to being fuzzy between multiple values which could be beneficial when multiple mixed decisions are required. We also found the "Learning to simulate complex physics with graph networks" really cool and relevant. It might be hard to implement it right away given the limited rebuttal time, but we will be happy to cite it in the final draft.

To R3:

We want to clarify a misunderstanding here. We do not detach the keys and queries (K_{sr} and Q_{sr}), else as you rightly mentioned, the FC layers preceding them would end up being random vectors. We detach the copy of features f_{sr} entering into the fully-connected layers which produce the keys and queries. This is also mentioned in lines $135-136$ and mathematically expressed in eqs 8 and 9. Note that gradients still backpropagate into f_{sr} via the yes-no responses: $V_{y,sr}$ and $V_{n,sr}$ since they do not received detached features. This way all layers have gradients backpropagated into them and no layer is left hanging.

While we have provided explanations for most quantities as they are introduced in section 3, we understand that it can be terse to parse the details in bits and pieces. We will remedy this with your suggestion and precede the detailed description with an overview paragraph explaining the architecture at a high level in the final draft. We will also expand our description of broader impact of our work and will be happy to take suggestions if you would like to offer some.

To R4:

The example in lines 19-22 is a motivational example for why continuous-valued decisions are required in decision making. We do not (over)-claim that our architecture will learn to make such human interpretable decisions.

In fact, there is no concrete way of forcing neural nets to learn human-interpretable representations on their own just by observing trajectory data without providing any additional human knowledge and we do not claim to do so (we explicitly mention this in lines 149-150). For instance, if quantities x and y are interpretable for humans, a neural net might easily learn other quantities, e.g. $x+y$ and $x-y$, which might be meaningless for humans but contain the same information as x and y. One can only check if our learnt decisions contain enough information about human-intepretable quantities and we have shown an instance of this in section 4.3 by predicting collisions solely from our learnt decisions (lines $245-253$).

Additionally, our architecture does allow providing pre-defined decisions which could come from human-knowledge and lets the model learn appropriate responses to them. This is not a trivial feature found in existing methods and is very useful in practical settings for debugging and for aligning the model to human expectations to some extent.