In this work, we tackle the task of question answering in 3D indoor environments. Previous methods are restricted to a pre-defined answer space. We propose Gen3DQA, an end-to-end transformer-based architecture to generate, rather than predict, natural answers for questions in 3D scenes. Our method directly optimizes the global semantics of the generated sentences via the language rewards.

After encoding the input scene and question into object proposals and questions embeddings, they are combined into one sequence and fed into a transformer encoder. The contextualized sequence is then fed into a transformer decoder to generate the answer.