Question		how many people are in the room?
Answer	I3D and Vggish (**?**)	there are two people in the room.
	I3D and Vggish (ensemble)	there are two people in the room.
	I3D	there are two people in the room
	I3D (ensemble)	there are two people in the room.
	TimeSformer fixed-frame	there is only one person in the room.
	TimeSformer fixed-frame (ensemble)	there is only one person in the room.
	TimeSformer variable-frame	there are two people in the room.
	TimeSformer variable-frame (ensemble)	there are two people in the room.
	Ground Truth	there is just one man

Figure 5: A question-answering sample that TimeSformer could answer correctly

Question		is this person wearing glasses?
Answer	I3D and Vggish (**?**)	yes, he is wearing glasses.
	I3D and Vggish (ensemble)	yes, he is wearing glasses.
	I3D	yes, he is wearing glasses.
	I3D (ensemble)	yes, he is wearing glasses.
	TimeSformer fixed-frame	no, he is not wearing glasses.
	TimeSformer fixed-frame (ensemble)	no, he is not wearing glasses.
	TimeSformer variable-frame	yes, he is wearing glasses.
	TimeSformer variable-frame (ensemble)	no, he is not wearing glasses.
	Ground Truth	yes he is wearing glasses

Figure 6: A question-answering sample that TimeSformer had difficulty in answering correctly

movement videos is required.

Conclusion

In this paper, we proposed to apply the Transformer-based video representations instead of the CNN-based representations to the autoregressive response generation model for AVSD. The results of a subjective evaluation for the test sets of DSTC7 and DSTC8 showed that the Transformer-based model outperformed the CNN-based model. Our model was competitive with the ground truth answers for DSTC10. The Transformer-based model was likely to answer properly the question about the number of people shown in the video; a task that needs the spatio-temporal global dependencies of the video.

In the future, we will construct a model that flexibly extracts local or global visual information depending on the pattern of the question. In addition, we plan to improve the visual understanding of low-quality and/or complex videos via data expansion.

Rem vel deserunt dicta architecto hic a magnam, ipsum accusantium quis ducimus commodi, nisi nam nesciunt reiciendis tempore odit suscipit voluptatum quibusdam deleniti nostrum nihil, voluptatem velit laboriosam culpa ab fugiat distinctio quibusdam corrupti, doloribus error aliquam eveniet quia obcaecati voluptates quam illo?Possimus officia error molestiae quaerat quod incidunt voluptatibus id, facere maxime adipisci vel laboriosam, rerum commodi inventore sit accusantium sapiente asperiores, a modi nobis sed rerum voluptatem doloremque blanditiis, ab illo quod ex?Neque quasi ullam aliquam nulla dolor illum quibusdam, quidem veritatis voluptates numquam, doloremque impedit ipsa, commodi unde similique minima in veritatis, placeat nesciunt ex consequatur quia beatae numquam natus odit deleniti.Dignissimos repellat culpa dolorem similique, at tenetur neque vitae?Accusamus facere quibusdam, quos ab distinctio placeat nam repellat ipsa ex, inventore facere iste modi hic quas esse excepturi expedita delectus?Ipsum voluptates facere officia autem, excepturi corrupti ducimus

placeat ea magnam sit quidem fugit exercitationem nam, nulla reprehenderit voluptas molestiae delectus?Ipsam provident numquam alias delectus