Reply

1. Is it possible to provide a visualization of multi model image problem, or schematic, for the readers in the introduction?

   Thanks for your question. We tried to draw, and finally found that the problem of multi-modal images is difficult to represent with visual images, but we added an explanation for this in the introduction.

2. How did you arrive at the number of convolutional layers? Did you go through a period of architecture optimization? e.g. using hyper tune?

   Thanks for your question. In order to get better discrimination effect, we optimized the network structure and performed a lot of training to compare and experiment. We tried a two-layer convolutional network for the first time, and found that the loss value of the discriminator soon stabilized, which to a certain extent reflects the weak discriminating effect, indicating that the network has room for improvement. In the end we got a 4-layer convolution discriminator, and many experiments show that it works best.

3. What is the role of data augmentation, if any, on the model performance? Is it possible to use synthetic data?

   Thanks for your question. Are you talking about data augmentation? The paper did not use this method, which may be a direction of our attention in the future.

4. The formatting of the equations seems to be wrong in places. Please check this e.g. in equation 1, the subscripts under argmax should be smaller than the main variables.

Thanks for your question. The expression of this formula means to extrapolate the next seven frames in the case of five inputs. The variables you mentioned can be understood as inputs.

5. Minor comments:

Thanks for your question. We have revised all questions and made relevant supplements in the paper.