SRVP: Strong Recollection Video Prediction Model Using Attention-Based Spatiotemporal Correlation Fusion

Supplementary Material

1. Qualitative Evaluations

1.1. Moving MNIST

Compared to RNN-based models, SRVP produces more accurate prediction results and demonstrates comparable performance to RNN-free models. While SimVP yields sharper outputs than RNN-based models, it exhibits inferior spatiotemporal representation capability compared to SRVP. On the whole, MIMO-VP achieves the highest image quality overall but tends to suffer from temporal inconsistency in long-term predictions.

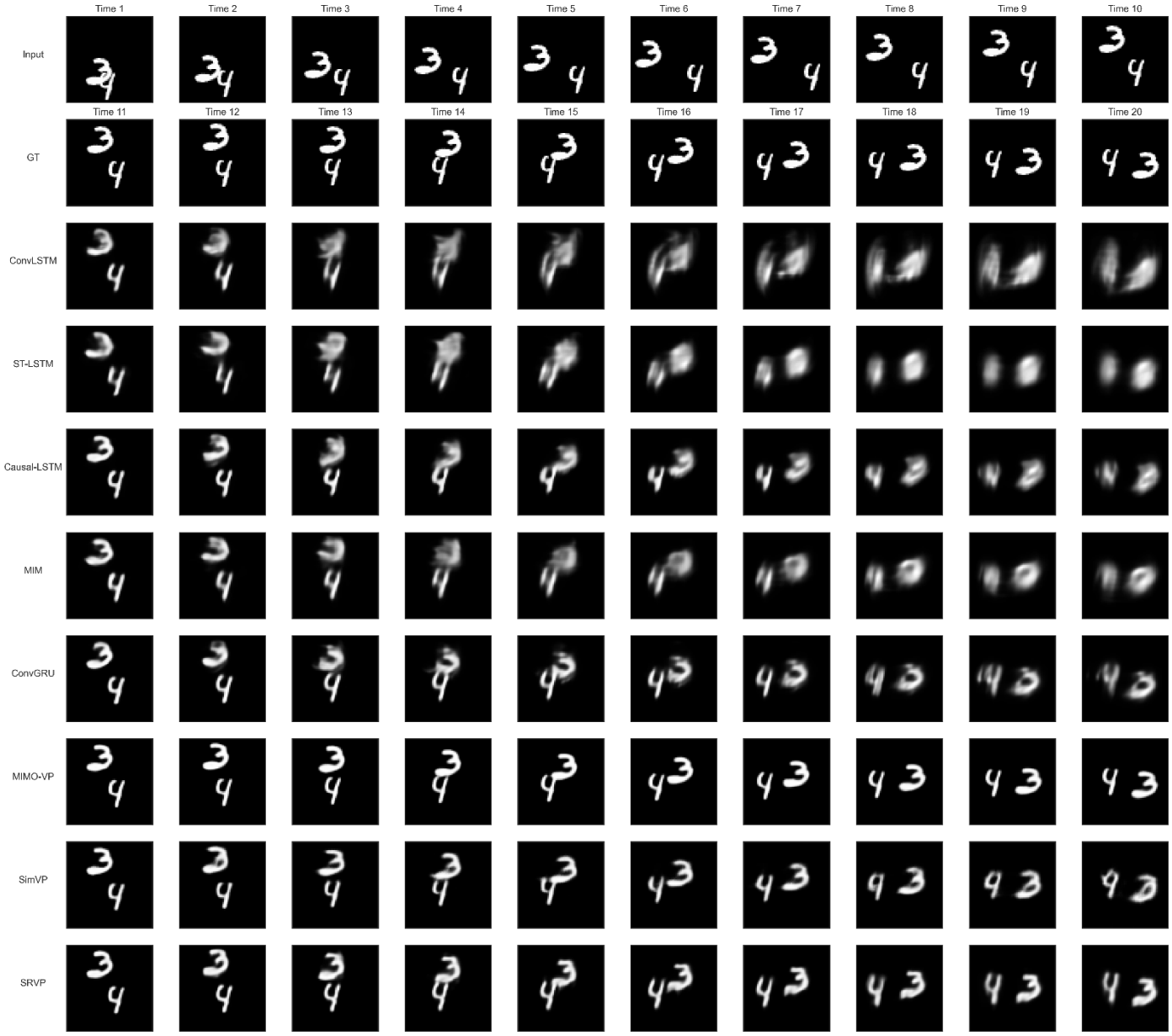

Figure 1. Prediction results on the Moving MNIST dataset (10 → 10).

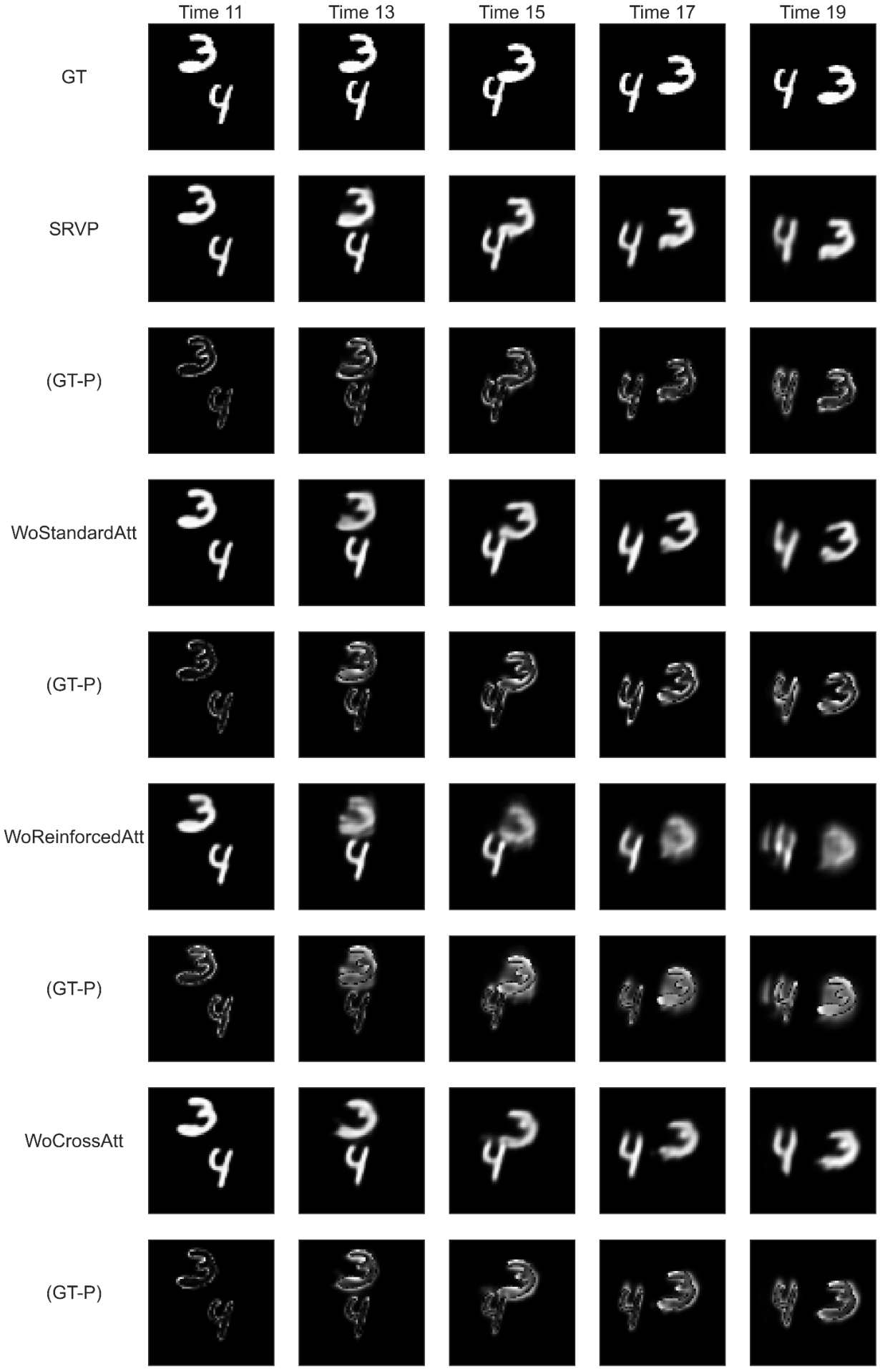

Figure 2. Ablation results on the Moving MNIST dataset (10 → 10). When the RFA module is removed from SRVP, the spatial information of moving objects is significantly degraded.

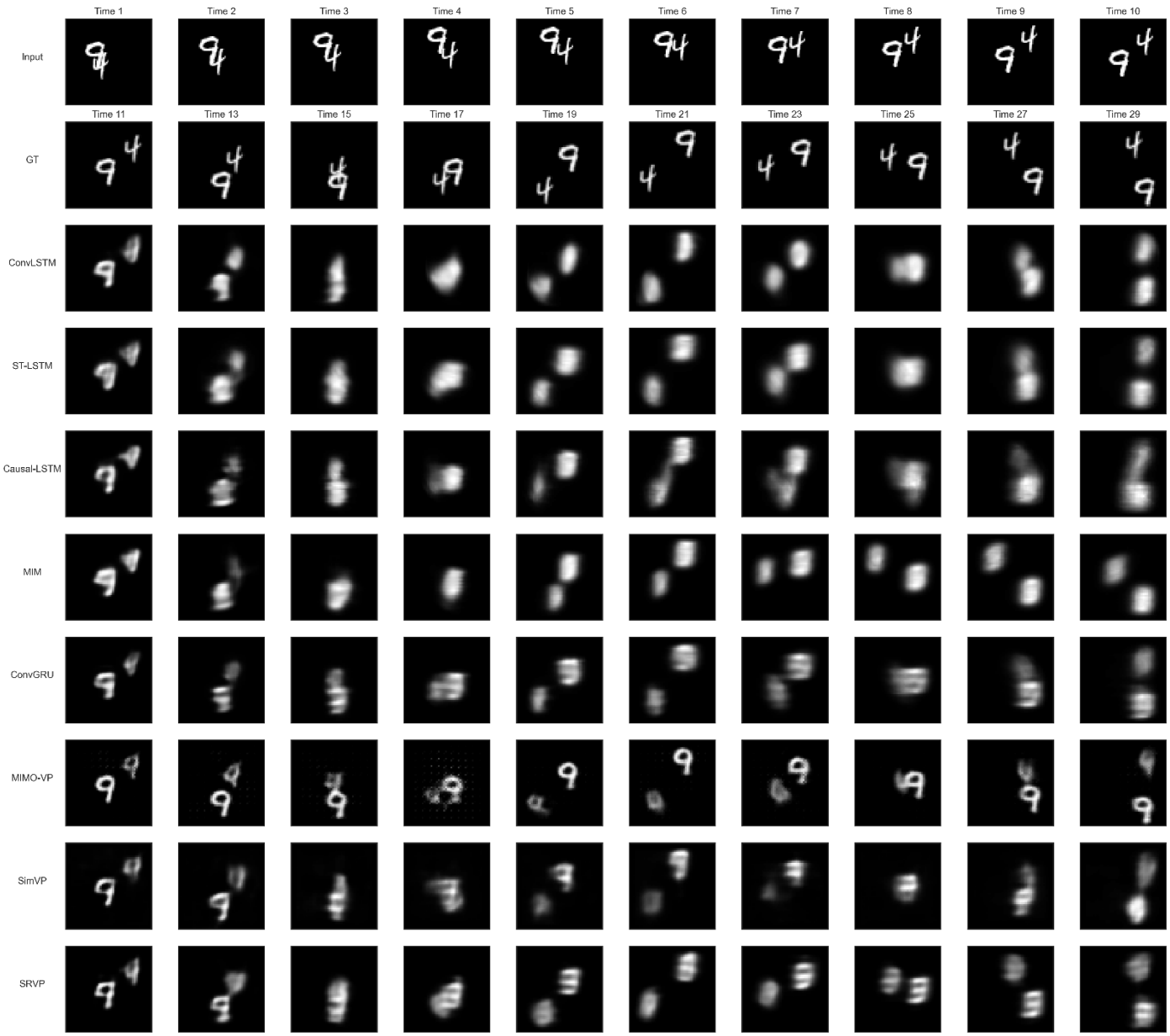

Figure 3. Prediction results on the Moving MNIST dataset (10 → 30). MIMO-VP shows dot artifacts in some time steps. This phenomenon is especially noticeable at Time 17.

1.2. KTH Action

RNN-based models tend to predict motion regions close to the average values, resulting in noticeable blurring artifacts. In contrast, RNN-free models exhibit fewer blurring artifacts in human body regions but struggle to capture temporal context, leading to inaccurate spatial positioning of limbs. Therefore, RNN-free models produce larger pixel-wise errors (MSE and PSNR) than RNN-based models and perform better when evaluated using structure-aware metrics (SSIM). This insight suggests potential directions for future research to improve the robustness and effectiveness of SRVP.

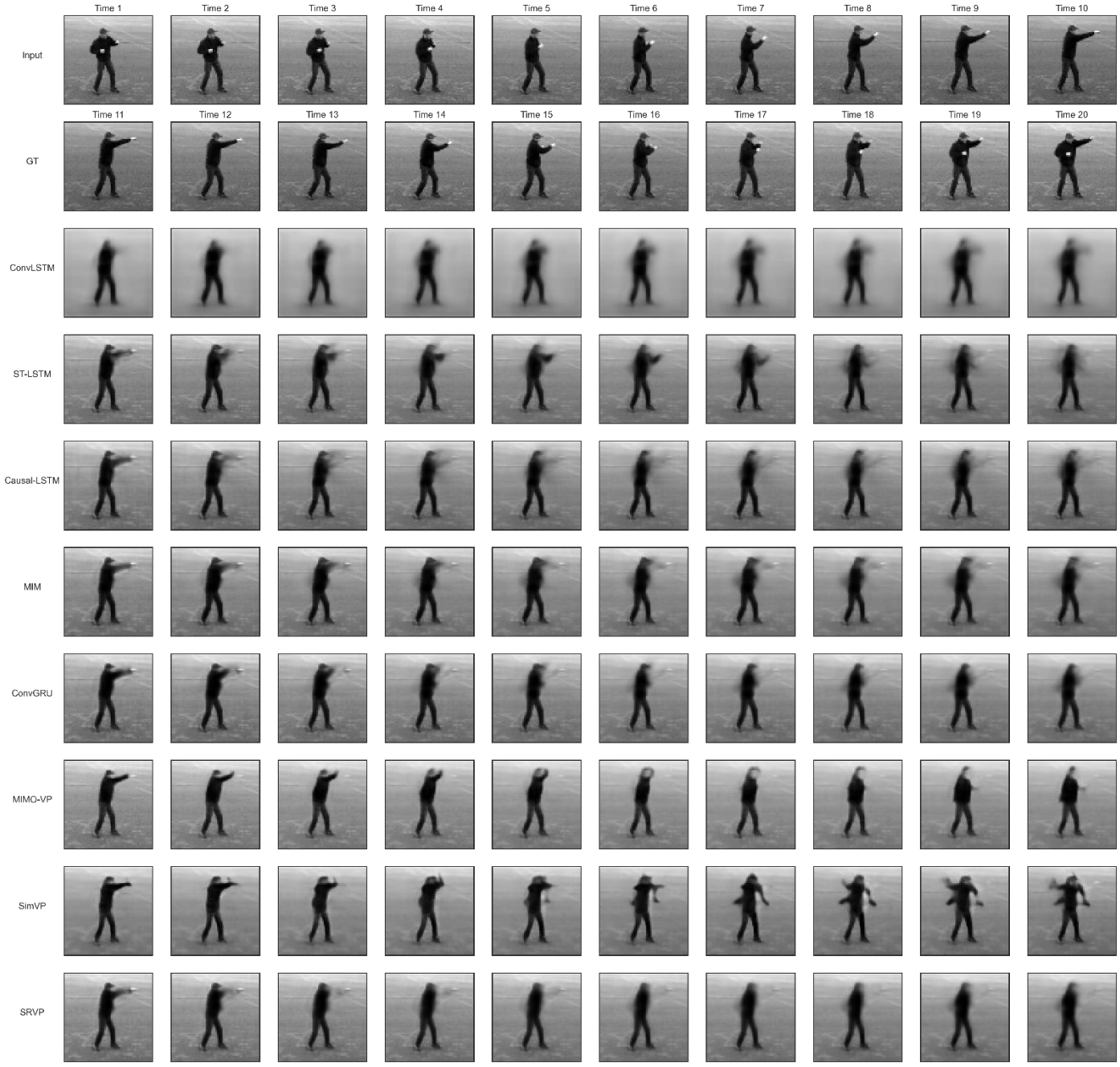

Figure 4. Prediction results on the KTH dataset (10 → 10).

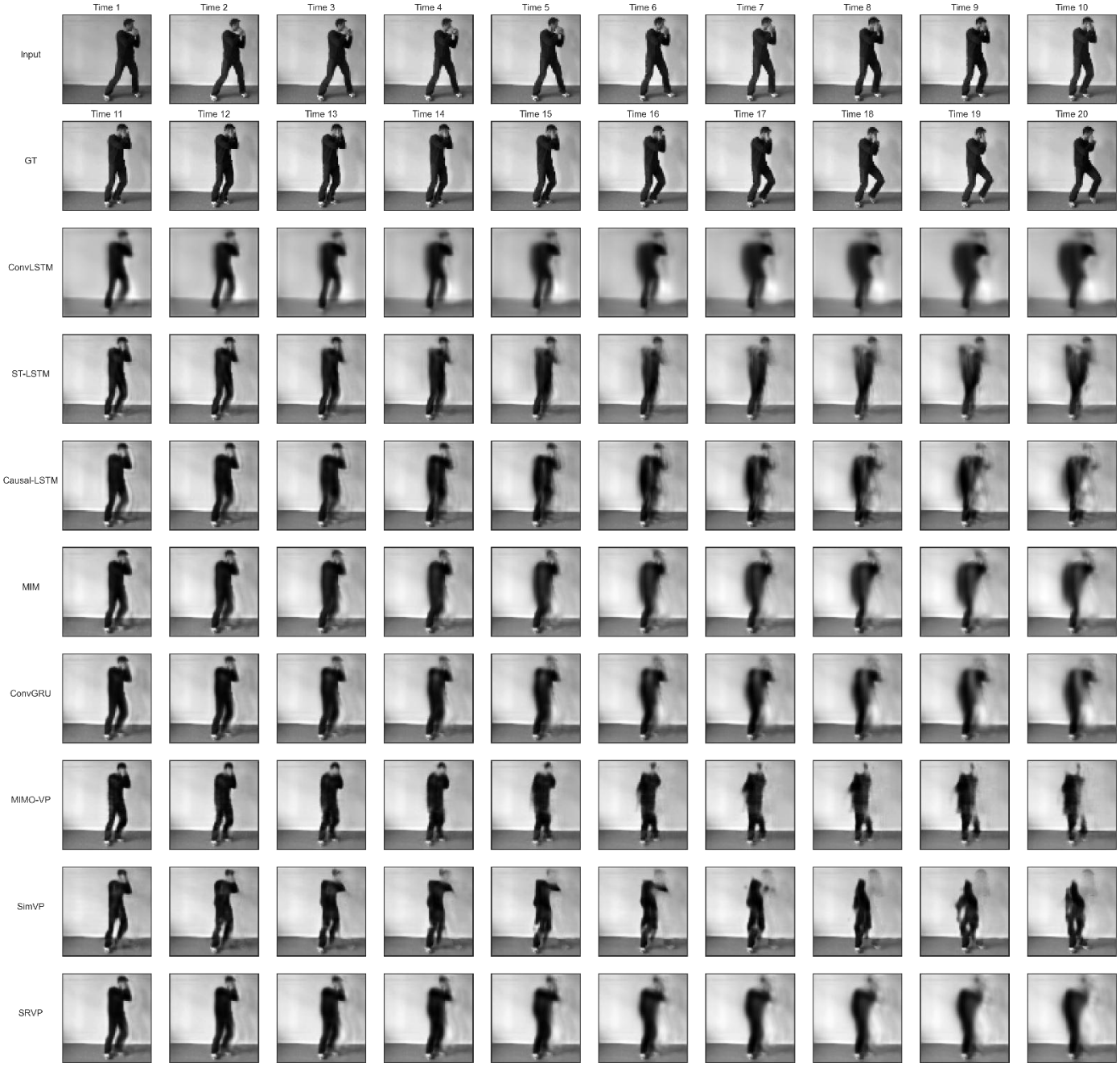

Figure 5. Prediction results on the KTH dataset (10 → 10).

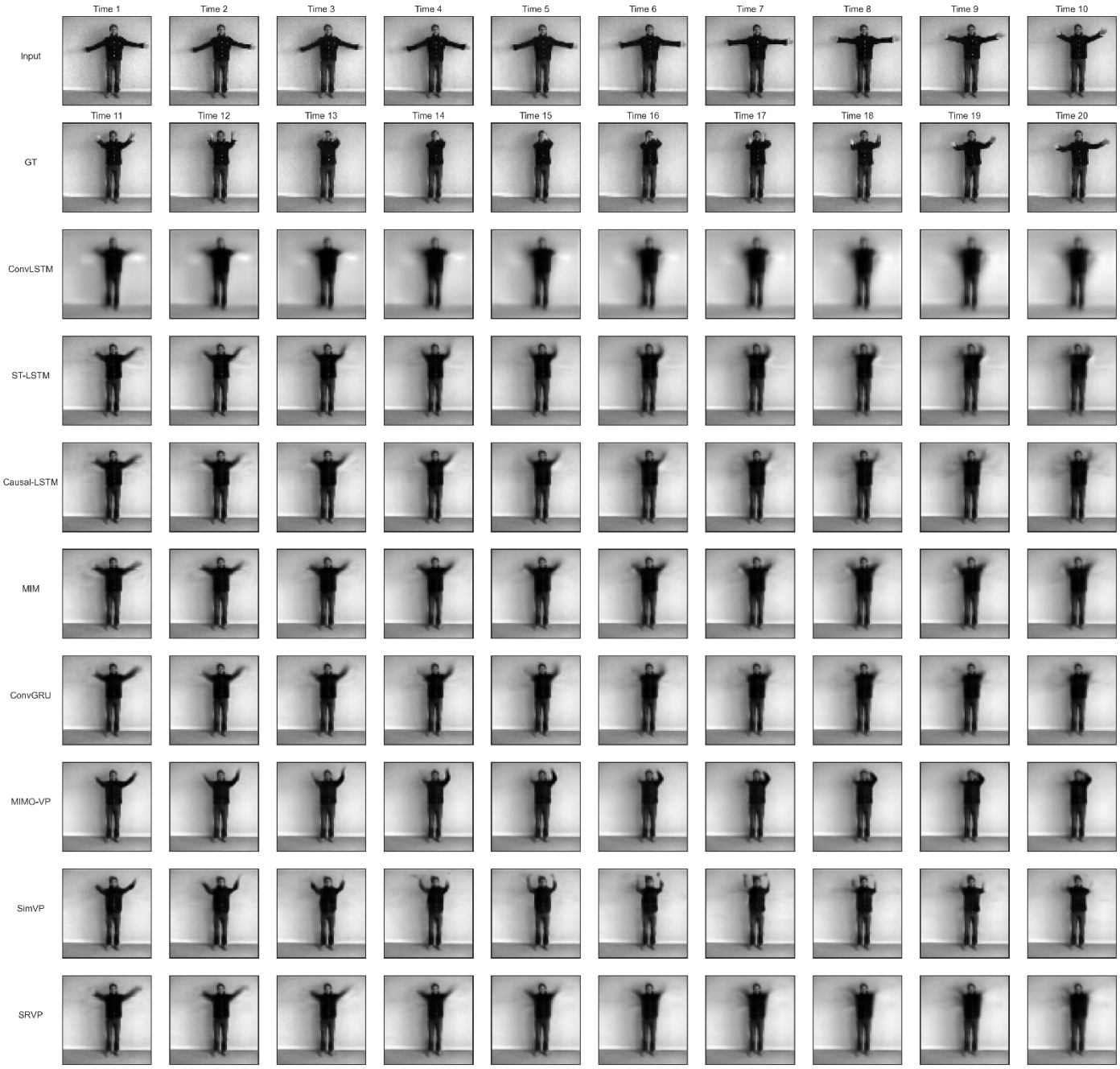

Figure 6. Prediction results on the KTH dataset (10 → 10).

1.3. Human3.6M

The prediction task becomes increasingly challenging as the image resolution grows. While all models struggle to accurately estimate human structures, SRVP shows a relatively better ability to preserve spatiotemporal context. A noteworthy observation is that RNN-free models produce considerable residuals not only in the positions of moving objects but also in the static background. RNN-free models heavily rely on learning spatial correlations, such as through convolutional operations. As image resolution increases and the visual content of videos becomes more complex, each receptive field or patch is required to process a significantly larger amount of information, which can limit the model's ability to capture fine-grained spatiotemporal details. This limitation can lead to a degradation in prediction accuracy. In contrast, RNN-based models use gating mechanisms to perform precise pixel-level analysis and more effectively capture spatiotemporal information. The prediction results presented below further support this insight.

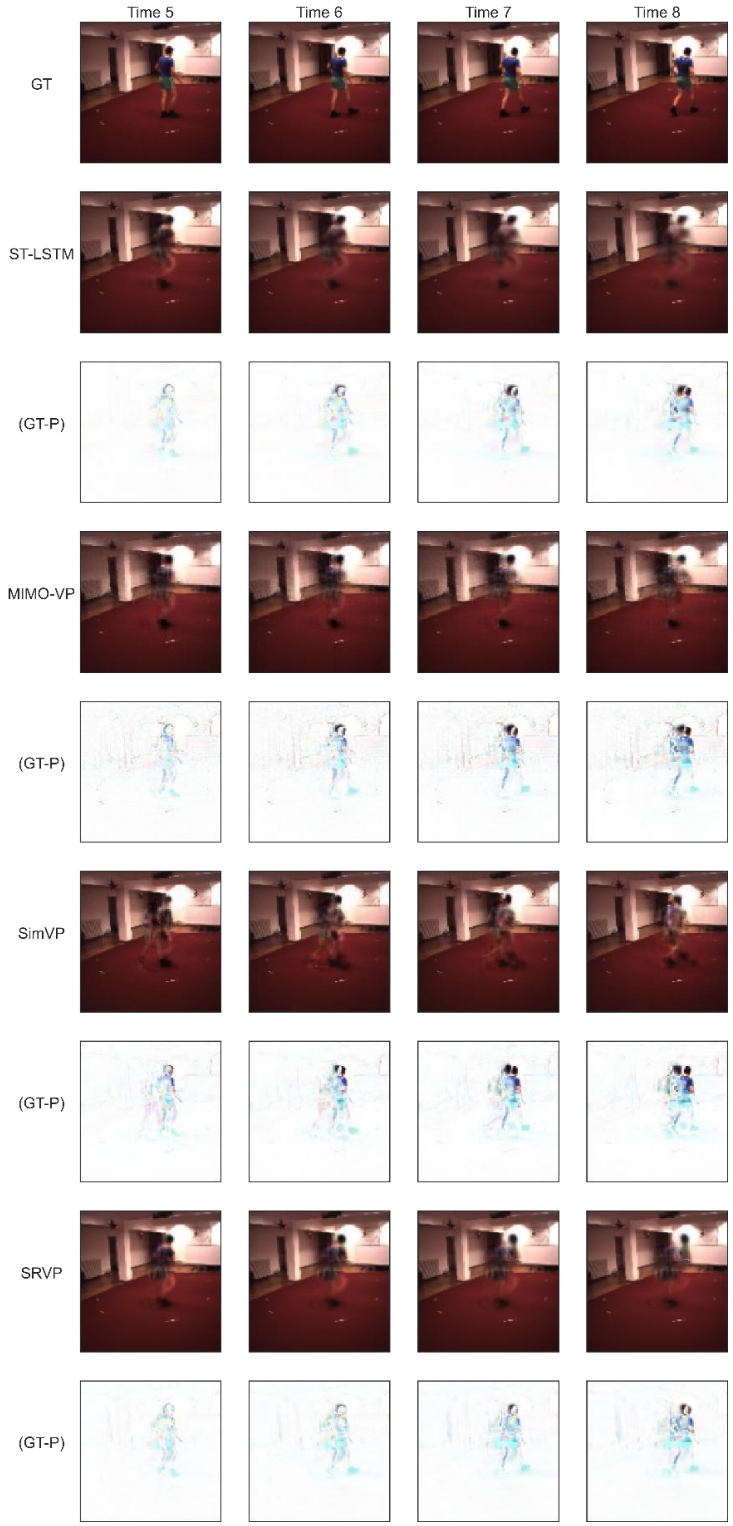

Figure 7. Prediction results on the Human3.6M (4 → 4) dataset.

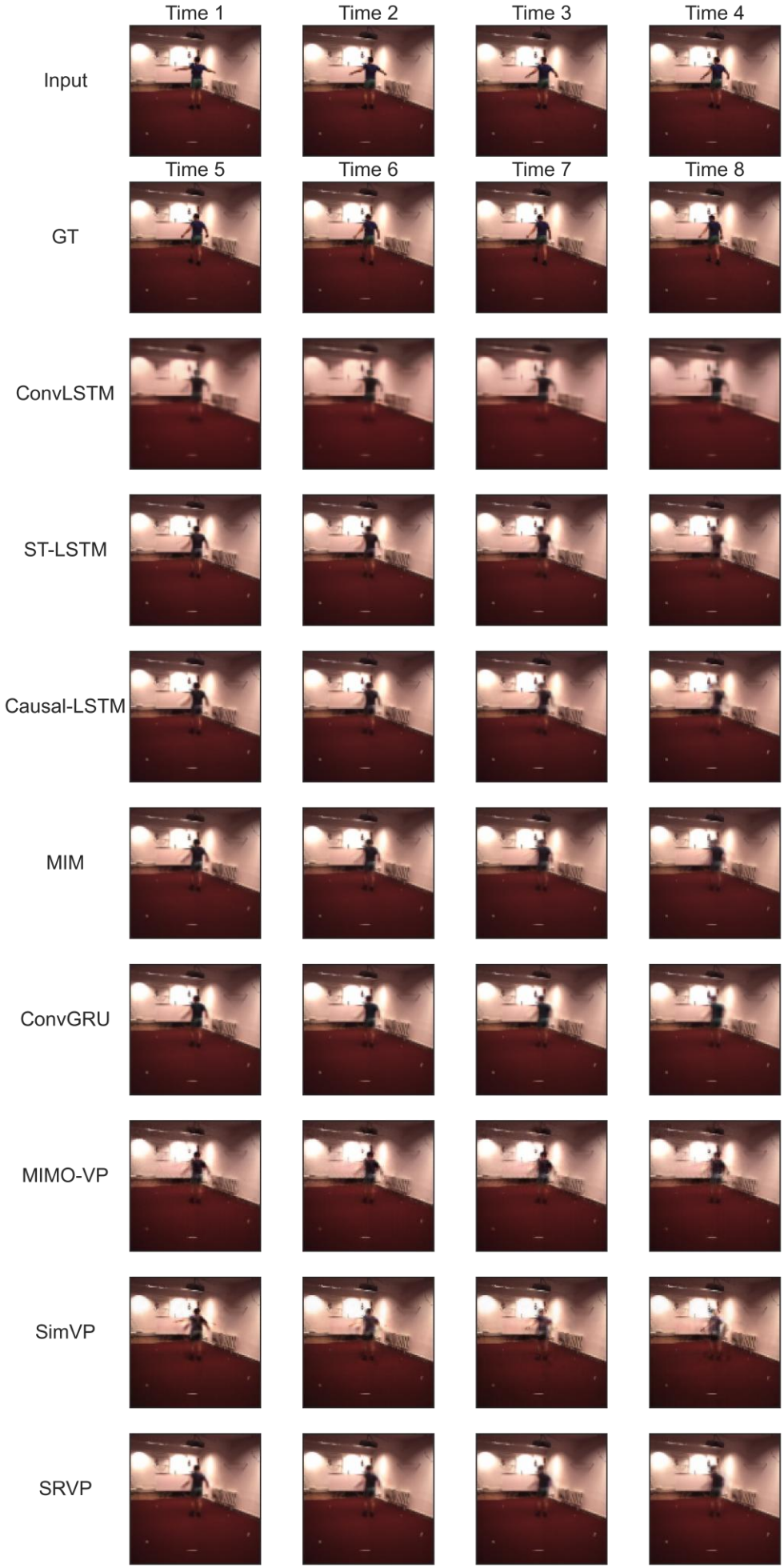

Figure 8. Prediction results on the Human3.6M dataset (4 → 4).

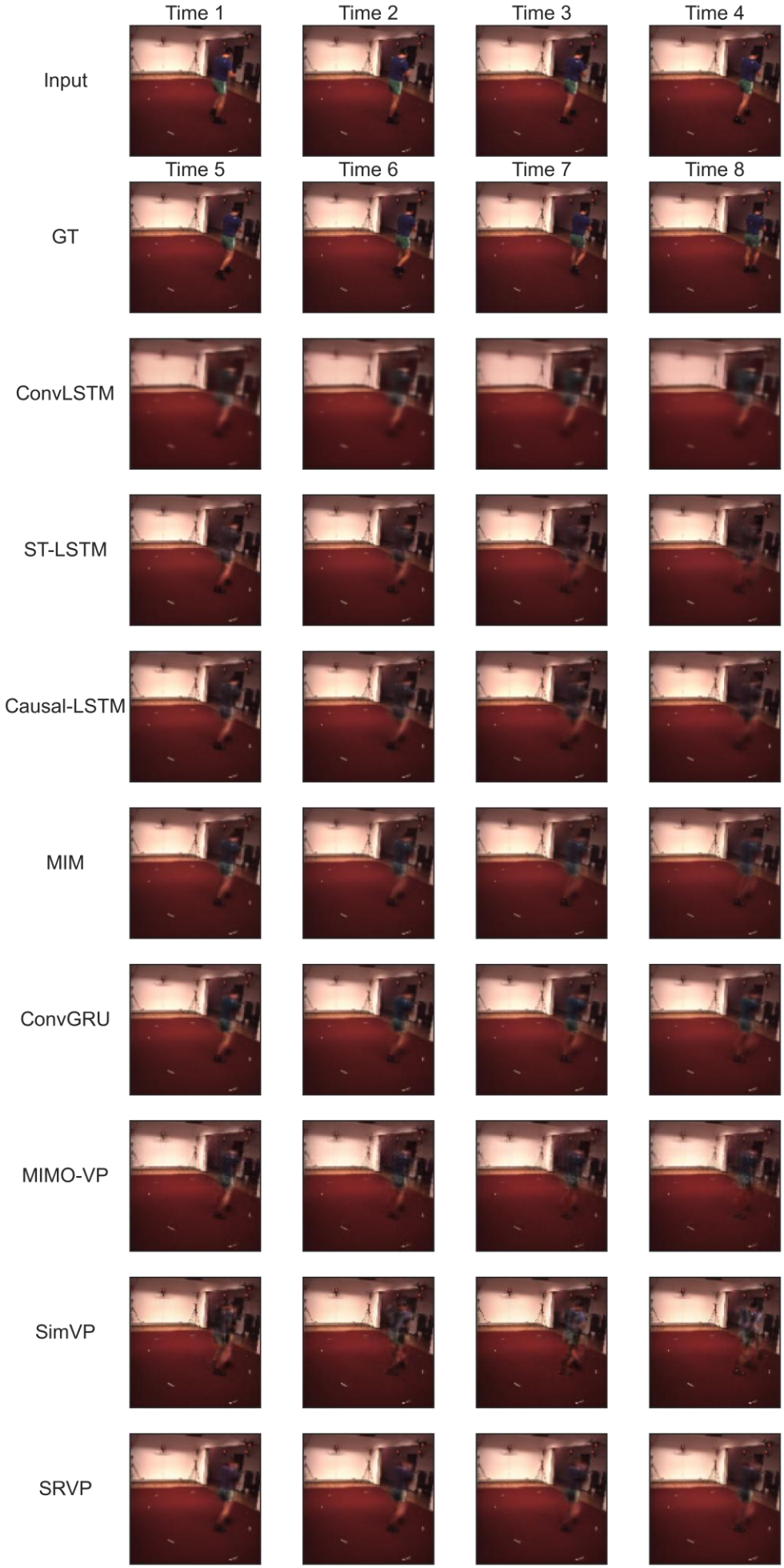

Figure 9. Prediction results on the Human3.6M dataset (4 → 4).