MIMO Is All You Need : A Strong Multi-In-Multi-Out Baseline for Video Prediction Supplementary Material

1 Evaluation Metrics

To evaluate the performance of our MIMO-VP against previous state-of-the-art works in the deterministic prediction domain, *i.e.*LMC-Memory (Lee et al. 2021), MotionRNN (Wu et al. 2021), PhyDNet (Guen and Thome 2020), ConvTT-LSTM (Su et al. 2020), CrevNet (Yu et al. 2019), MIM (Wang et al. 2019) and ConvLSTM (Xingjian et al. 2015), several metrics are used in our experiment, including Mean Squared Error (MSE), Mean Absolute Error (MAE), the Structural Similarity (SSIM), LPIPS (Zhang et al. 2018) and Peak Signal to Noise Ratio (PSNR). All these metrics are averaged for all predicted frames. Lower MSE, MAE, LPIPS, or higher SSIM, PSNR indicates better performance. In addition, for the Weather dataset, we also take the Critical Success Index (CSI) into account. The CSI score is a significant evaluation metric in the precipitation nowcasting domain, which is calculated as:

$$CSI = \frac{hits}{hits + misses + falsealarms}, \tag{1}$$

where *hits*, *misses* and *falsealarms* denote the number of *true positive*, *false positive* and *false negative*, respectively. The higher the CSI score, the better the performance.

2 Implementation Details

More implementation details are given in this section. ReduceLROnPlateau is adopted to reduce the learning rate with a 0.5 decay factor when loss stops decreasing. For the Moving MNIST dataset, we use 6 transformer blocks in the encoder and 10 transformer blocks in the decoder. While for the other three datasets, we use both 6 blocks in the encoder and decoder modules. The model channel is set to be $C = 128$. We use 8 heads in our 2DMHA module. The batch size is set to be 32 for the Moving MNIST dataset and 8 for other datasets.

3 Additional Experimental Results

3.1 How to make the best design?

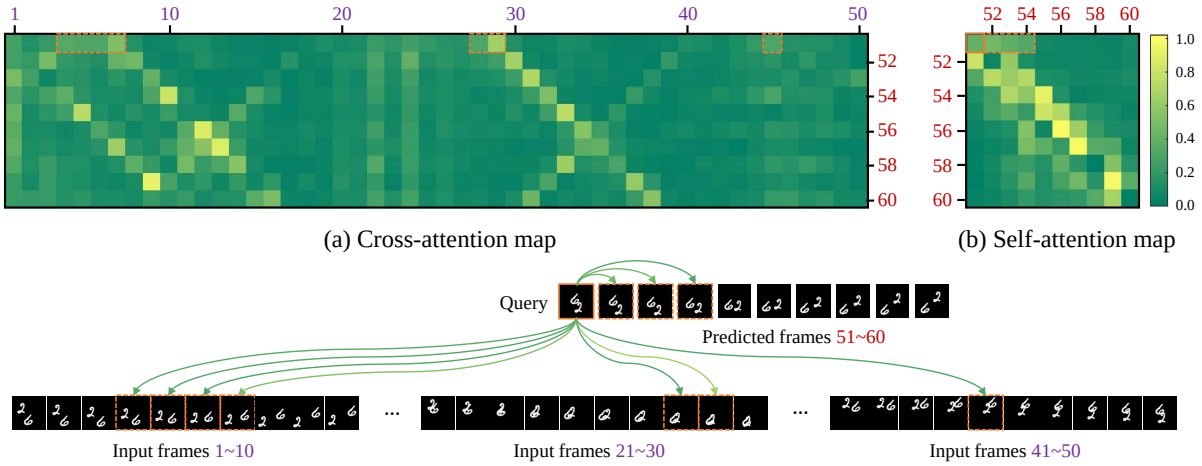
We report the experimental results regarding the number of transformer blocks and the number of heads in 2DMHA in Tab. 1. Generally, a deeper model with more layers of the transformer will bring higher performance. For the trade-off of prediction accuracy and model efficiency, we use '6-10' model on Moving MNIST, and '6-6' model on the other datasets. We note that the results of multiple heads' attention are better than single heads in our experiment because multiple heads could effectively capture complicated long-term temporal cues for video prediction. Therefore, we chose 8 heads in our MIMO-VP.

Table 1: Quantitative comparison of different designs of MIMO-VP on Moving MNIST dataset. Model 'a-b' denotes MIMO-VP with 'a' encoder blocks and 'b' decoder blocks.

Models	MSE ↓	MAE ↑	SSIM ↑
4-8	19.3	55.4	0.960
6-10	17.7	51.6	0.964
8-12	17.6	51.3	0.964
1 head	19.1	55.0	0.960
4 heads	18.0	52.4	0.963
8 heads	17.7	51.6	0.964
16 heads	18.1	52.7	0.963

3.2 More illustration about the superiority of MIMO

To investigate how the MIMO is capable of capturing the long-term relationships of sequence and preserving the future frame dependency, we conduct experiments on the Moving mnist dataset by predicting 10 future frames conditioned on 50 observed frames. The cross-attention map in Figure 1 (a) shows that the predicted frames have stronger attention with the distant input frames (small index of frames), which indicates that MIMO-VP can leverage the distant historical information of sequence based on the 2DMHA mechanism to facilitate future frame generation. For example, Figure 1 (c) shows that the input frames 4, 5, 6, 7, 28, 29 and 45 are helpful for predicting frame 51. In addition, the self-attention map in Figure 1 (b) shows that each predicted frame has high attention scores with its neighbouring frames, indicating their dependency. This phenomenon verifies the benefit of MIMO-VP in preserving the dependency among the future frames for accurate prediction.



(a) Cross-attention map

(b) Self-attention map

(c) Illustration of the predicted frame 51

Figure 1: The attention maps of MIMO-VP, at 4-th layer with the sum of 8 heads: (a) cross-attention map shows that the generated frames have high attention scores with the distant historical frames; (b) self-attention map indicates the dependency among the future frames; (c) an illustration of dependency between the predicted frame 51 and the other frames.

3.3 More Visualizations

We give in Figure 2, Figure 3 and Figure 4 additional visualizations completing Figure 5, 6 and 3 in submission, respectively. We see that the frames generated by MIMO-VP are clearer than those generated by other comparison methods in Figure 2. In Figure 3, our MIMO-VP achieves higher prediction accuracy than MIM, MotionRNN and PhyDNet, especially in the region of high-intensity radar echos. In Figure 4, MIMO-VP consistently generates accurate and sharp frames, while the number four becomes distorted when they are predicted by PhyDNet and MotionRNN.

We also provide the long-term prediction comparison results on the Moving mnist dataset, as shown in Figure 5. We see that numbers four and zero in the top sample of PhyDNet and MotionRNN become seriously distorted after timestep 22 when these two numbers become overlap. Similarly, numbers six and eight in the bottom sample of PhyDNet and MotionRNN become distorted and blurry after timestep 14. Surprisingly, our MIMO-VP can consistently predict accurate and clear frames. These results prove the effectiveness of MIMO-VP in long-term prediction.

3.4 Demo

We put a demo in the attachment, which shows the dynamic visualization of different predictions.

References

Guen, V. L.; and Thome, N. 2020. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11474–11484.

Lee, S.; Kim, H. G.; Choi, D. H.; Kim, H.-I.; and Ro, Y. M. 2021. Video Prediction Recalling Long-term Motion Context via Memory Alignment Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3054–3063.

Su, J.; Byeon, W.; Kossaifi, J.; Huang, F.; Kautz, J.; and Anandkumar, A. 2020. Convolutional Tensor-Train LSTM for Spatio-Temporal Learning. *Advances in Neural Information Processing Systems*, 33.

Wang, Y.; Zhang, J.; Zhu, H.; Long, M.; Wang, J.; and Yu, P. S. 2019. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9154–9162.

Wu, H.; Yao, Z.; Wang, J.; and Long, M. 2021. MotionRNN: A flexible model for video prediction with spacetime-varying motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15435–15444.

Xingjian, S.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, 802–810.

Yu, W.; Lu, Y.; Easterbrook, S.; and Fidler, S. 2019. CrevNet: Conditionally Reversible Video Prediction. *arXiv preprint arXiv:1910.11577*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

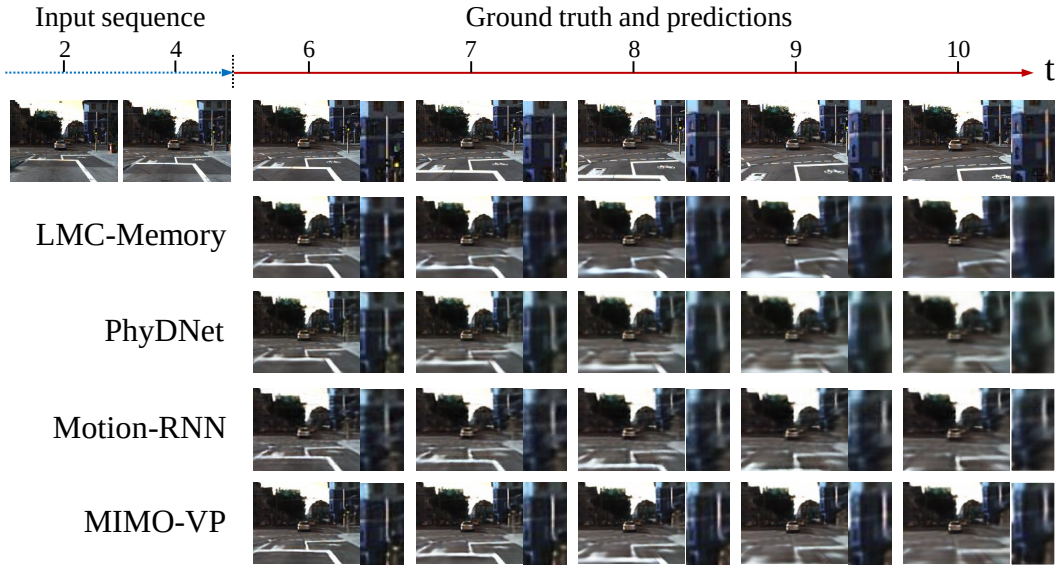


Figure 2: Prediction examples on KITTI dataset. We predict 5 future frames conditioned on 5 observed frames.

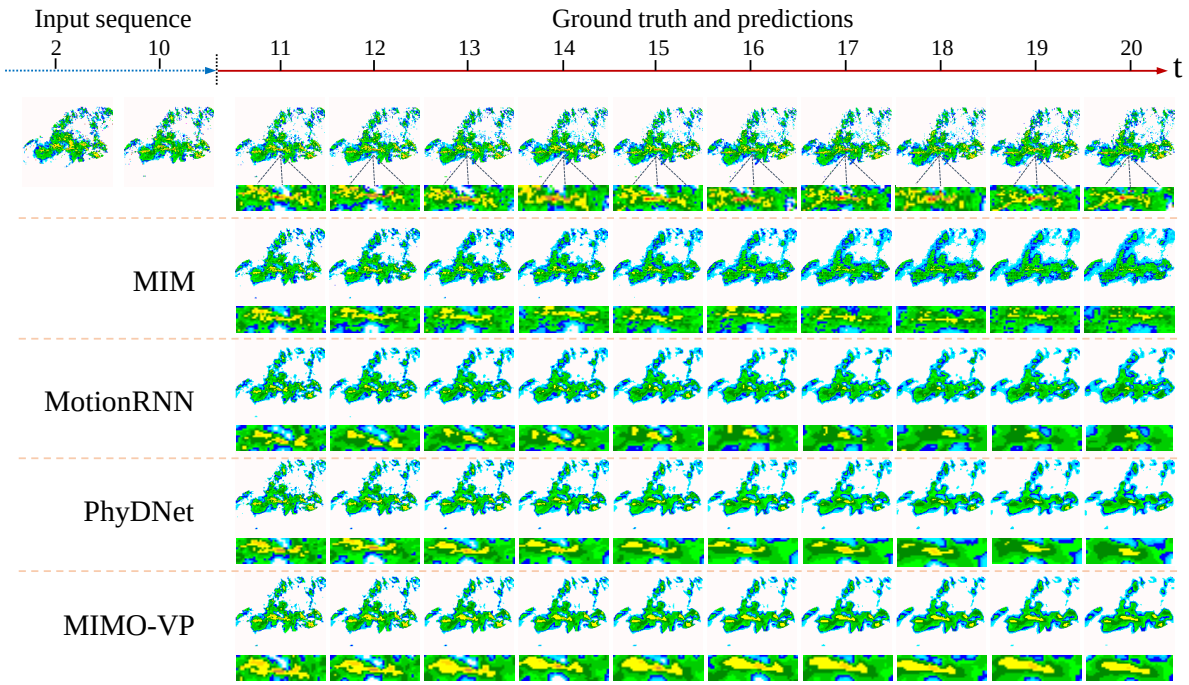


Figure 3: Prediction examples on Weather dataset. We predict 10 future frames conditioned on 10 observed frames.

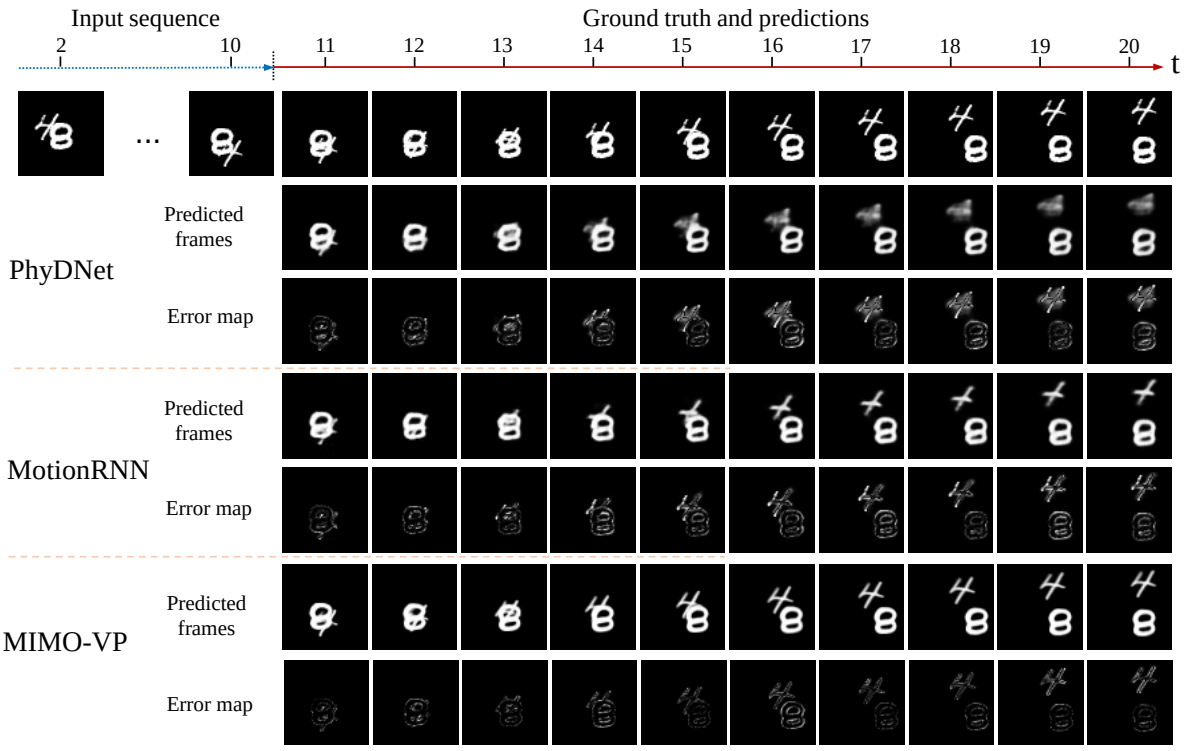


Figure 4: Prediction examples on Moving mnist dataset. We predict 10 future frames conditioned on 10 observed frames.

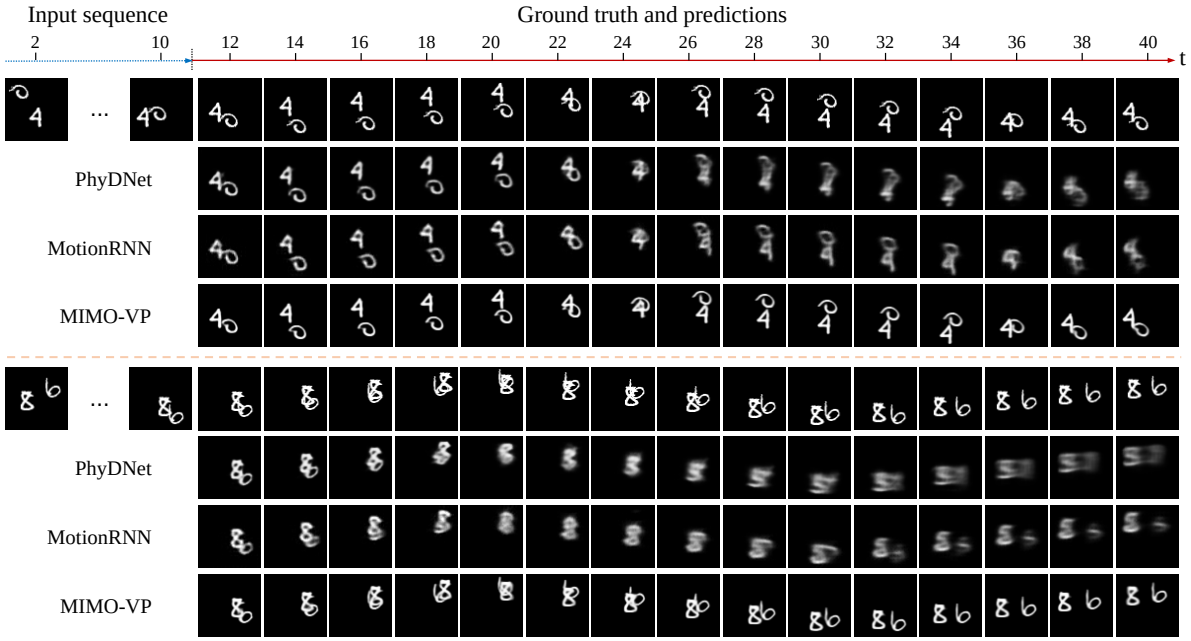


Figure 5: Prediction examples on Moving mnist dataset. We predict 30 future frames conditioned on 10 observed frames. Our MIMO-VP can correctly predict the movement of numbers zero and four in the top sample and numbers six and eight in the bottom sample, while these numbers become distorted when they are predicted by PhyDNet(Guen and Thome 2020) and MotionRNN (Wu et al. 2021).