A EXPERIMENTAL SETTINGS AND DATASETS

Datasets We quantitatively evaluate our model on the following datasets for both synthetic and real-world scenarios:

- **Moving MNIST** Srivastava et al. (2015) is a synthetic dataset consisting of two digits independently moving within the 64 × 64 grid and bouncing off the boundary. It is a standard benchmark in spatiotemporal predictive learning.
- **Human 3.6M** Ionescu et al. (2013) is a large-scale 3D human motion capture dataset for fitness, close human interactions, and self-contact. This dataset contains 3.6 million 3D human poses and corresponding images, 11 professional actors (6 male, five female), and 17 scenarios (discussion, smoking, taking photos, talking on the phone, etc.).
- **Weather Benchmark** Rasp et al. (2020) This dataset contains various types of climatic data from 1979 to 2018. The raw data is regrind to low resolutions, we here choose $5.625°$ (32×64 grid points) resolution for our data. Since the complete data is very large and includes massive climatic attributes like geopotential, temperature, and other variables, we specifically chose the global temperature prediction task to evaluate our model.
- **Caltech Pedestrian** is a driving dataset focusing on detecting pedestrians. It consists of approximately 10 hours of 640 × 480 videos taken from vehicles driving through regular traffic in an urban environment. We follow the same protocol of PredNet Lotter et al. (2017) and CrevNet Yu et al. (2019) for pre-processing, training, and evaluation.
- **KTH** Schuldt et al. (2004) contains 25 individuals performing six types of actions. Following Villegas et al. (2017); Wang et al. (2018b), we use persons 1-16 for training and 17-25 for testing. Models are trained to predict the next 20 frames from the previous 10 observations.

We summarize the statistics of the above datasets in Table 5, including the number of training samples N_{train} and the number of testing samples N_{test}.

Table 5: The statistics of datasets. The training or testing set has N_{train} or N_{test} samples, composed by T or T' images with the shape (C, H, W).

	N_{train}	N_{test}	(C, H, W)	T	T'
MMNIST	10,000	10,000	(1, 64, 64)	10	10
Human 3.6M	73,404	8,582	(3, 256, 256)	4	4
WeatherBench	2,167	706	(1, 32, 64)	12	12
Kitti&Caltech	3,160	3,095	(3, 128, 160)	10	1
KTH	4,940	3,030	(1, 128, 128)	10	20

Baselines We choose the following baselines for comparison: (i) Recurrent-based methods including ConvLSTM Shi et al. (2015), PredRNN Wang et al. (2017), PredRNN++ Wang et al. (2018a), MIM Wang et al. (2019a), E3D-LSTM Wang et al. (2018b), and PredRNNv2 Wang et al. (2021); (ii) Recurrent-free methods including SimVP Gao et al. (2022a), TAU Tan et al. (2023a), Uniformer Li et al. (2022), MLP-Mixer Tolstikhin et al. (2021), and ConvNeXt Liu et al. (2022).

Measurement We employ Mean Squared Error (MSE), Mean Absolute Error (MAE), Structure Similarity Index Measure (SSIM), and Peak Signal to Noise Ratio (PSNR) to evaluate the quality of predictions. MSE and MAE estimate the absolute pixel-wise errors, SSIM measures the similarity of structural information within the spatial neighborhoods, and PSNR is an expression for the ratio between the maximum possible power of a signal and the power of distorted noise. LPIPS Zhang et al. (2018) is a perceptual similarity metric that computes the distance between two images' feature representations in a pre-trained deep network.

Implementation details We implement the proposed method with the Pytorch framework and conduct experiments on a single NVIDIA-V100 GPU. The AdamW optimizer is utilized with a learning rate of 0.01 and a weight decay of 0.05.

B DISCUSSION ABOUT THE TEMPORAL STRIDE

The choice of the temporal stride Δt plays a crucial role in navigating the trade-off between performance and efficiency, thereby impacting the performance of USTEP in various spatiotemporal prediction tasks. In contrast, ΔT is set to be the same as T for all datasets to capture the macro-temporal scale dependencies. Table 6 outlines the selected values for Δt across different datasets, providing insights into the alignment of the model's focus with the inherent characteristics.

Table 6: The choices of Δt and ΔT for different datasets.

	MMNIST	Human 3.6M	WeatherBench	Caltech	KTH
T	10	4	12	10	10
T'	10	4	12	1	20
Δt	5	2	4	5	10
ΔT	10	4	12	10	10

The value of Δt is chosen to be half of T for MMNIST, reflecting a balanced approach to incorporating both micro- and macro-temporal scale information. This approach is mirrored in Caltech dataset as well, with Δt being half of T to ensure the synthesis of local and global perspectives. For Human 3.6M, a smaller Δt is selected to give more weight to micro-temporal scale dependencies, given the dataset's nuanced temporal dynamics. Similarly, in WeatherBench, a Δt of 4 is chosen to provide a balanced view of the temporal sequence, catering to the dataset's diverse temporal patterns. For the KTH dataset, Δt equals T as the dataset is relatively simple, allowing the model to harness the whole input temporal context within only one micro-temporal segment to generate coherent and plausible future frames.

C VISUALIZATION

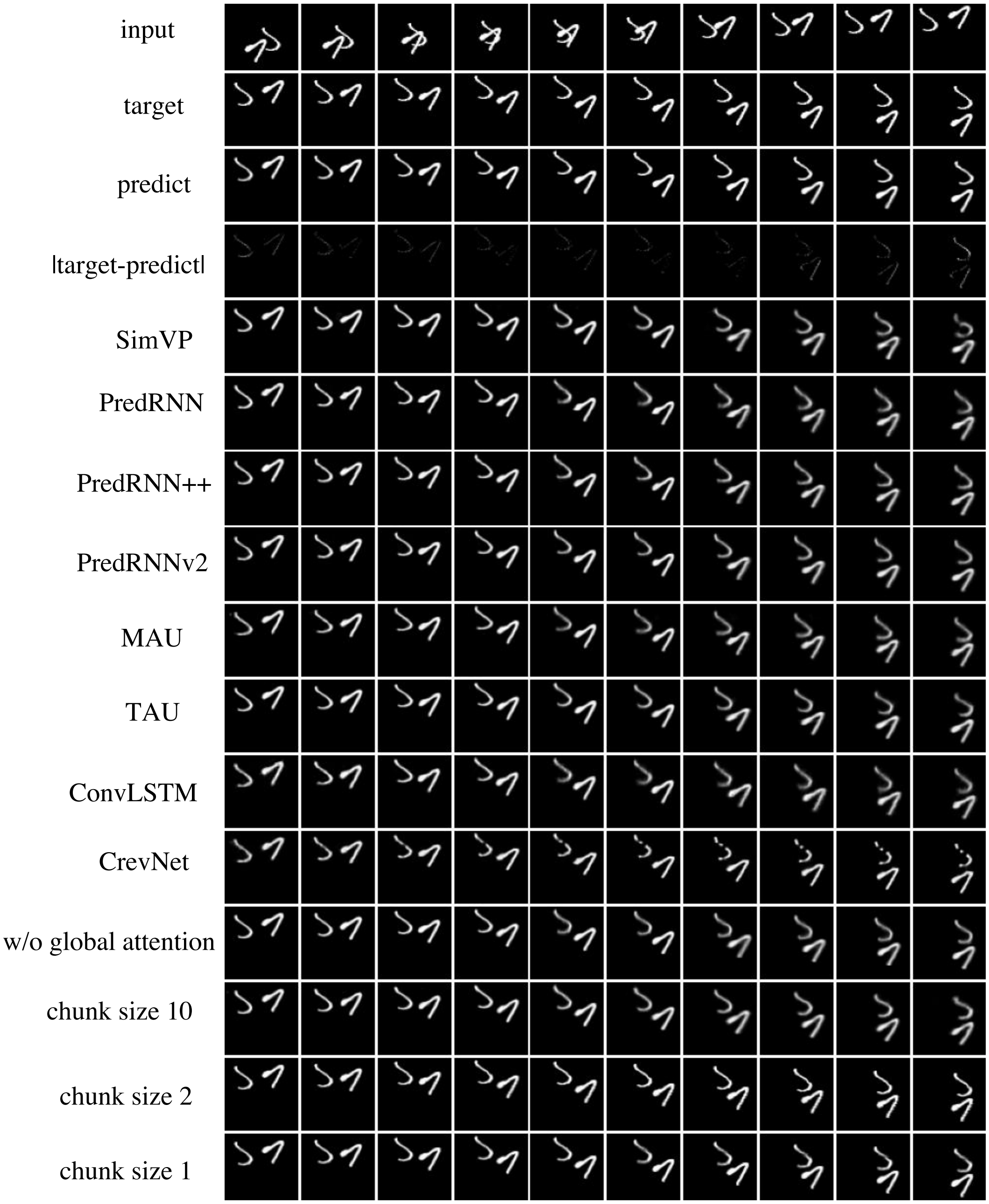

Figure 5: The qualitative visualization on Moving MNIST.

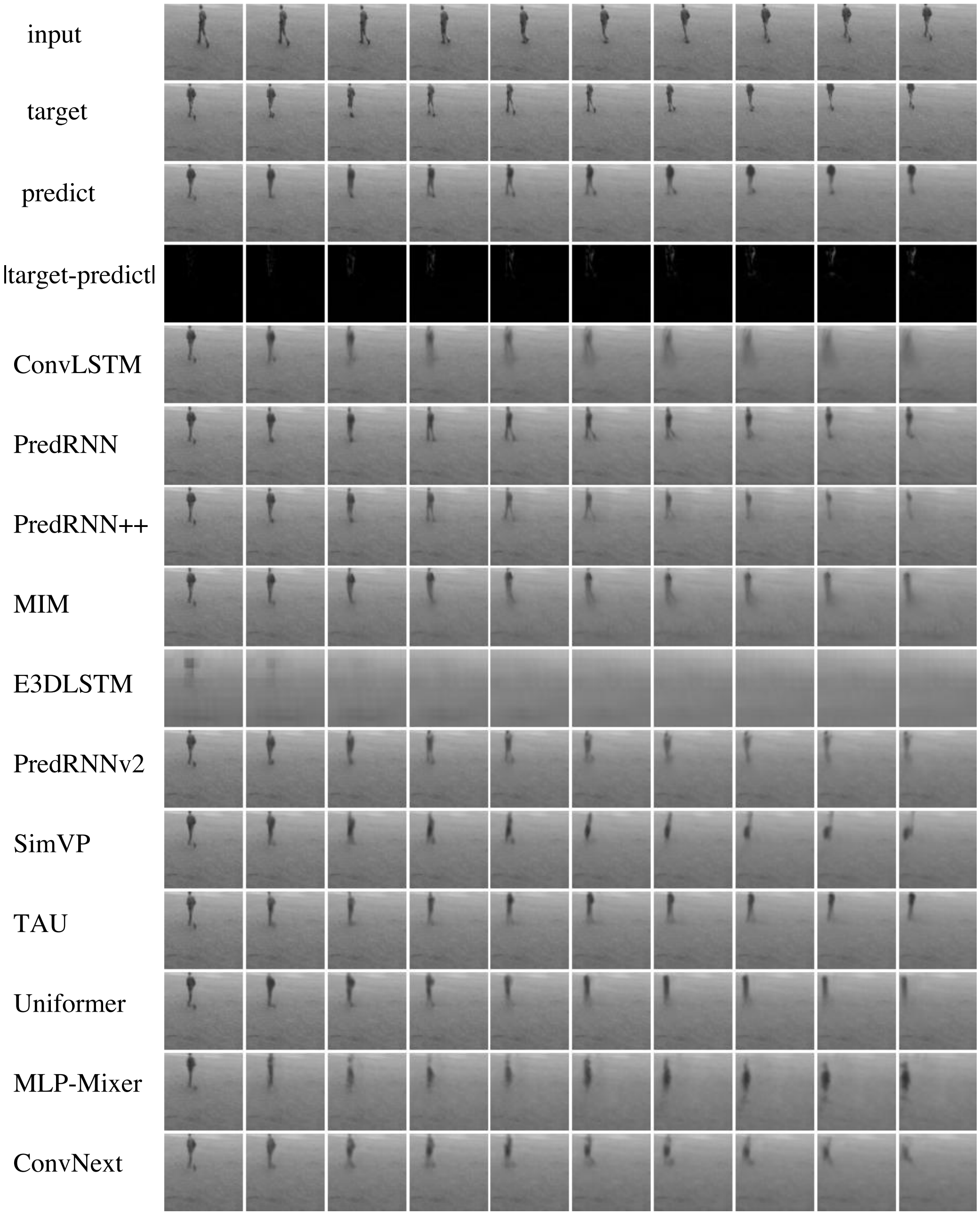

Figure 6: The qualitative visualization on KTH dataset. The target and predicted sequence is the range of {2,4,...,20}

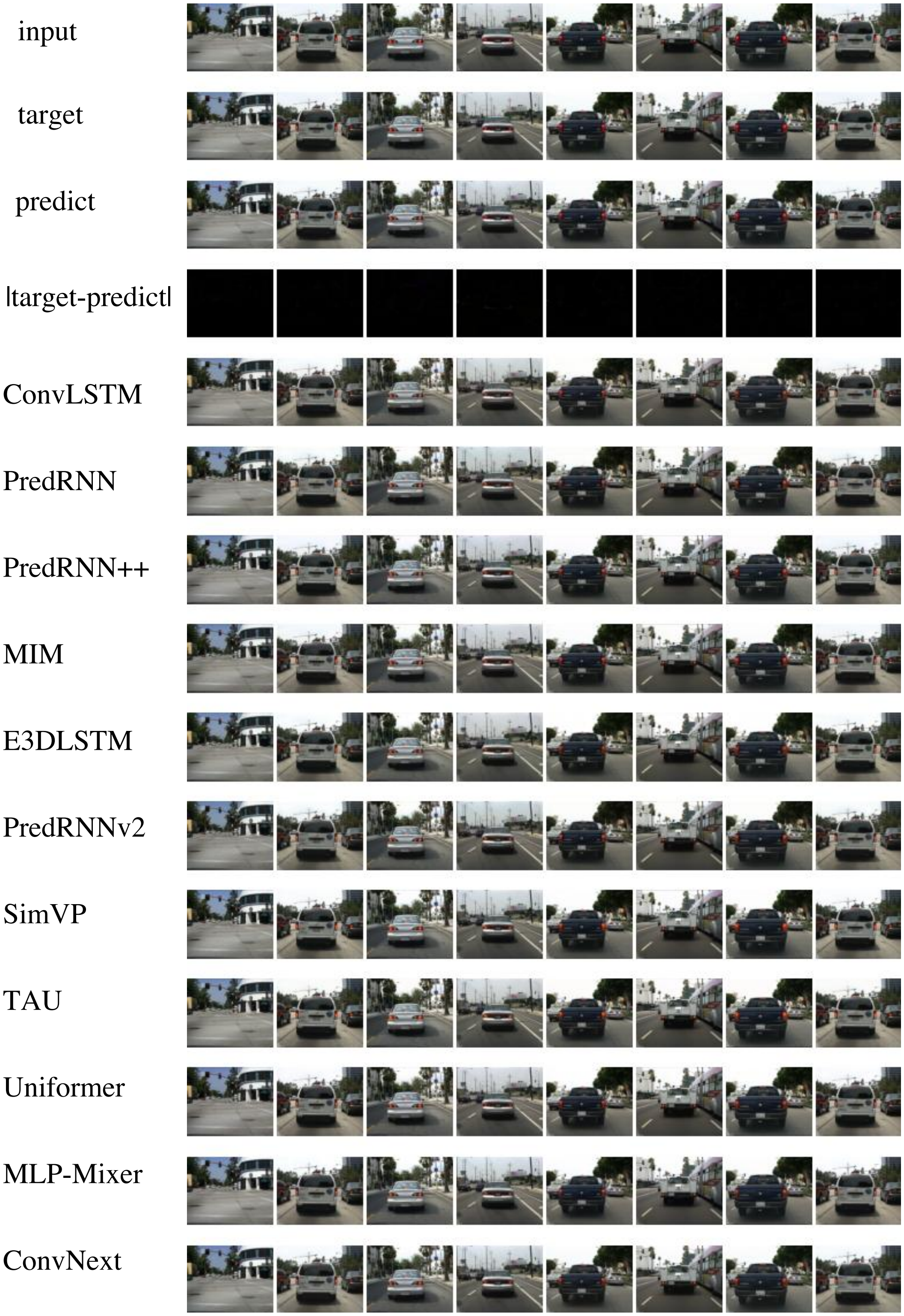

Figure 7: The qualitative visualization on Caltech Pedestrian dataset.

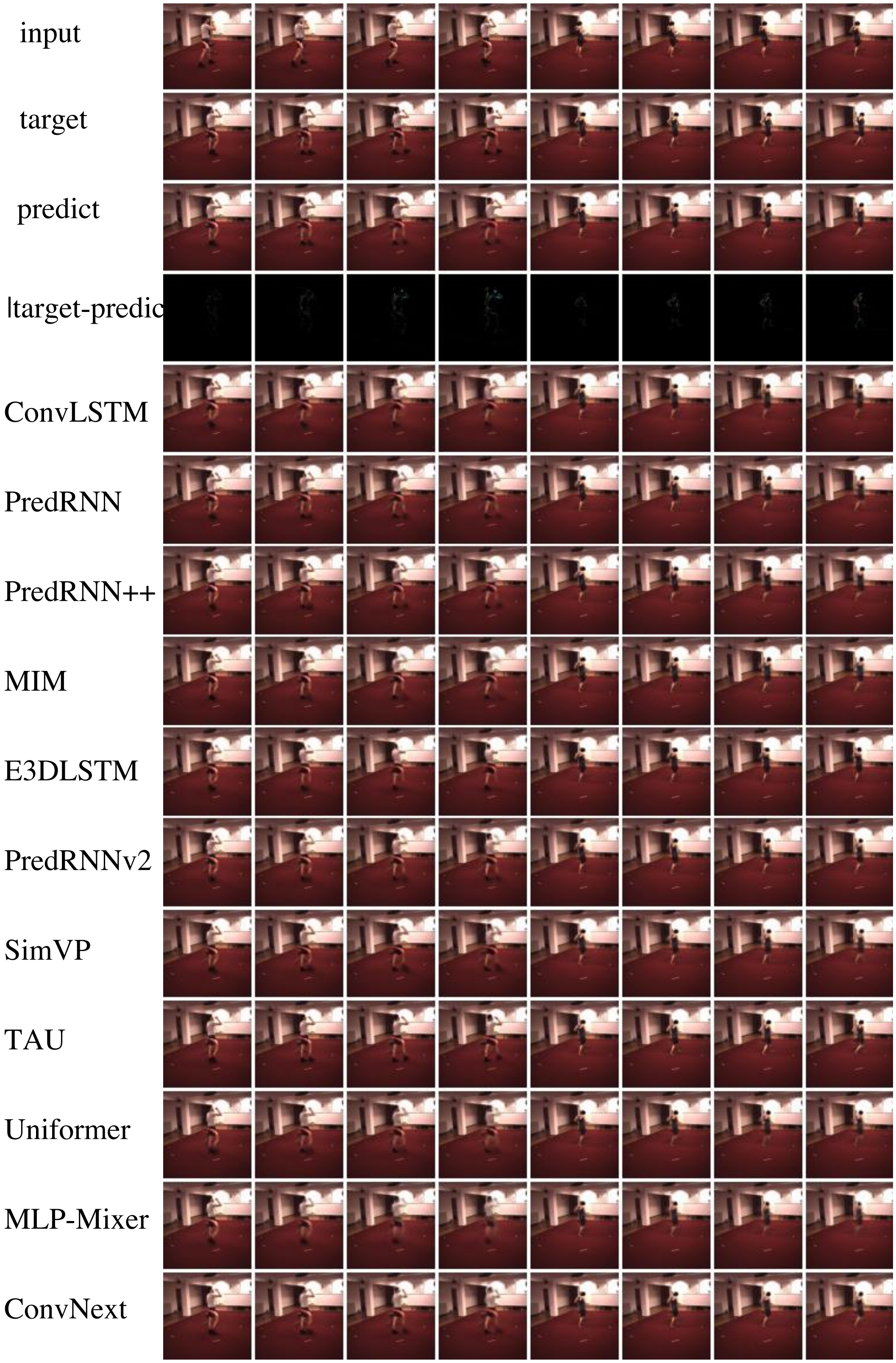

Figure 8: The qualitative visualization on Human 3.6M.