Zero-shot Video Prediction with Moving MNIST variants on OpenSTL

Moving MNIST data

In the canonical Moving MNIST the dataset consists of short video clips where MNIST digits move across a 64x64 canvas.

  - Sequence:10 input ->10 output frames.
  - Default training and validation sizes: 10,000. generated dynamically during training.
  - Testing size: 10,000. pre-generated in this study to maintain the consistency of variant experiments.

- Frame:
  - Completely black canvas (or a variant of a random CIFAR-10 image background). Each frame has either 1 channel (grayscale) or 3 channels (RGB), with pixel values scaled to [0,1].
  - Add 2 original MNIST digits (28x28 size), may overlap.
- Predictable motion patterns:
  - During training, new sequences are generated by randomly sampling starting positions (x,y) and directions θ
  - Digits follow straight line motion with a fixed step size=0.1, and when they reach the image boundary their trajectory is reflected elastically.
- Data augmentation:
  - OpenSTL provides: resizing, cropping, flips, and adding more MNIST numbers to the image.
  - We extend further movie modification: modifying step sizes (tested), varying step sizes and directions with frames, and adding position-dependent sampling.

Model Setup

- Default SimVP-gSTA and convLSTM-L configurations in OpenSTL.
- Recall Train : val : test -> 10,000:10,000:10,000
- Behavior:
  - SimVP: 200 epochs, best at 179 epochs, ~4 min/epoch
  - convLSTM: early stop at 20 epochs, best model at 15 epochs, ~40 min/epoch

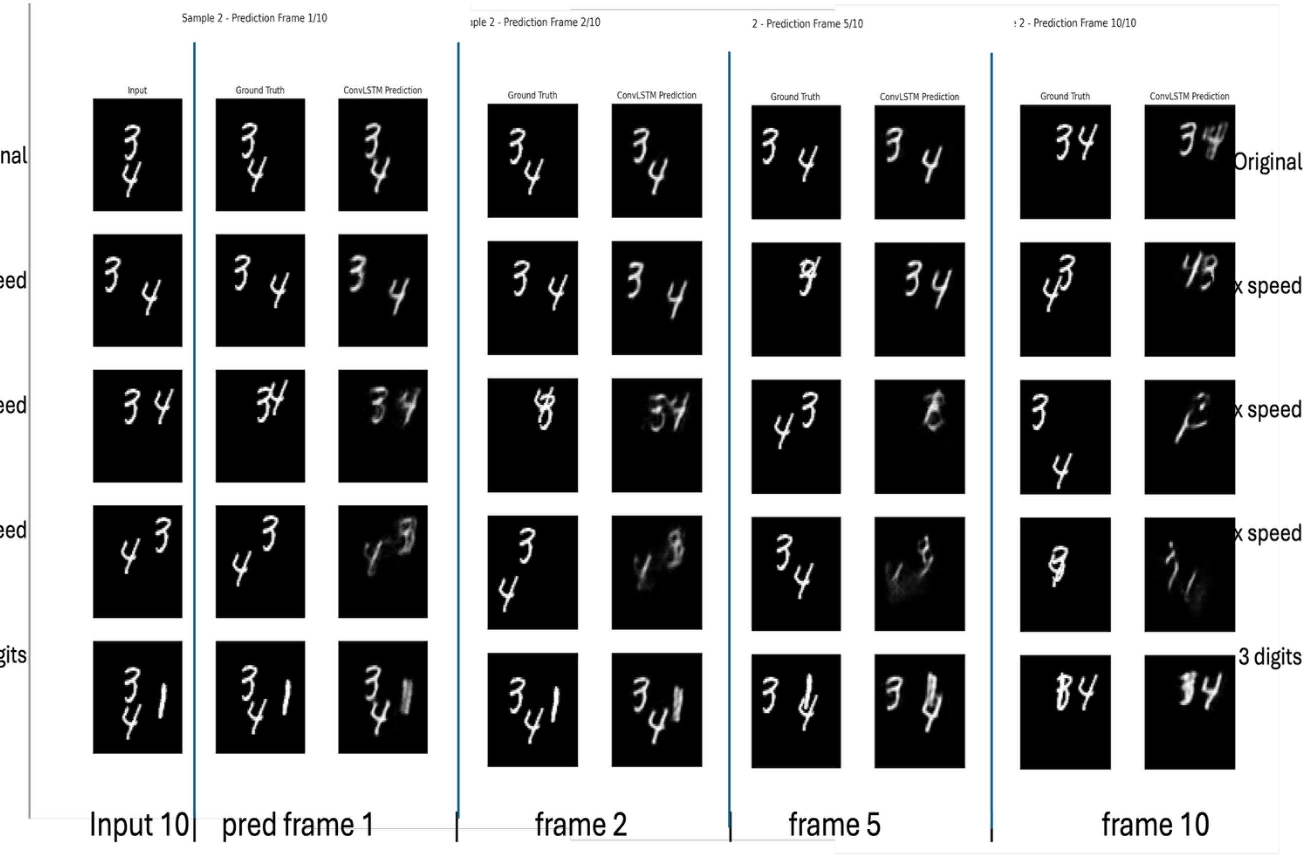
  -> Similar runtime: ~13 hr.
- Test Metrics

Metric	SimVP	ConvLSTM
MAE	104.62	126.12
MSE	39.24	64.37

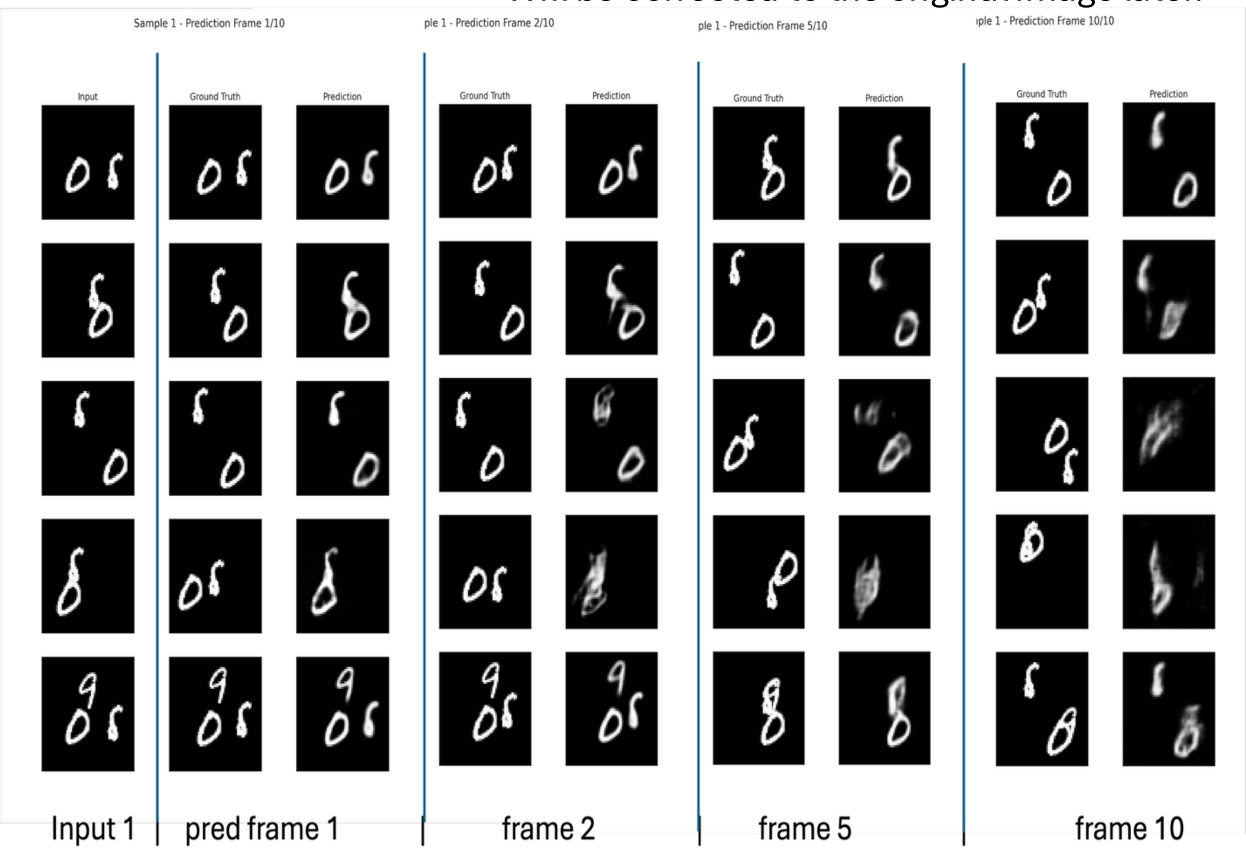
Brief Comparison between models

The digits are thicker for simvp because we enhanced the image to check for the presence of gray pixels in the background.
Will be corrected to the original image later.

convLSTM



simVP

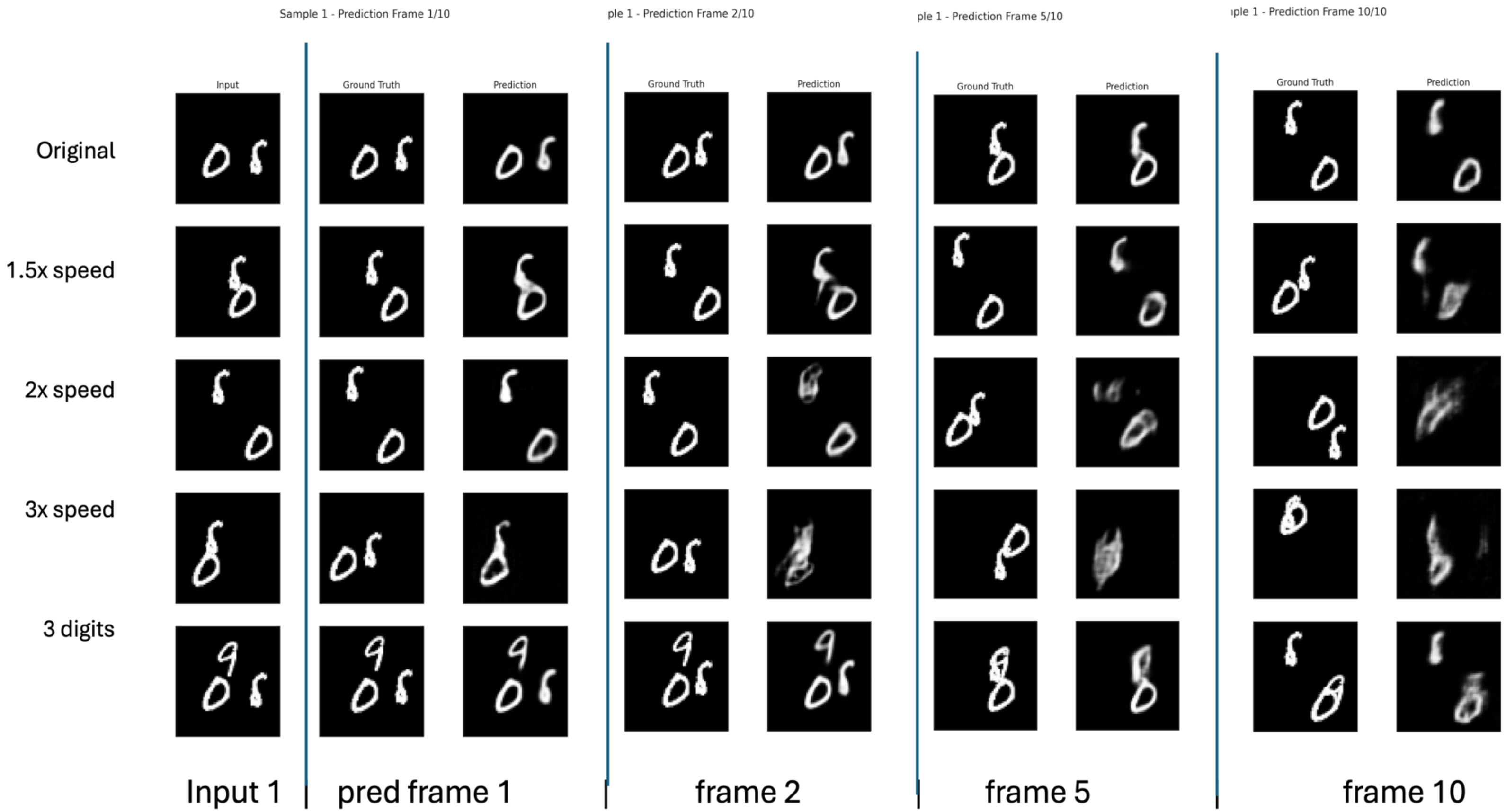


Zero-shot on Variants – SimVP-gSTA

Train on the original Moving MNIST dataset, test the data variants with the best model.

Findings: (refer to the movies at next slides)

- At 2× speed, predicted motion speed lower → reflects the original speed in training
- Faster motion → predictions blur quickly
- Easy to get strong blur at overlaps (two digits collide), especially with 3 digits



Sample Movies from SimVP-gSTA

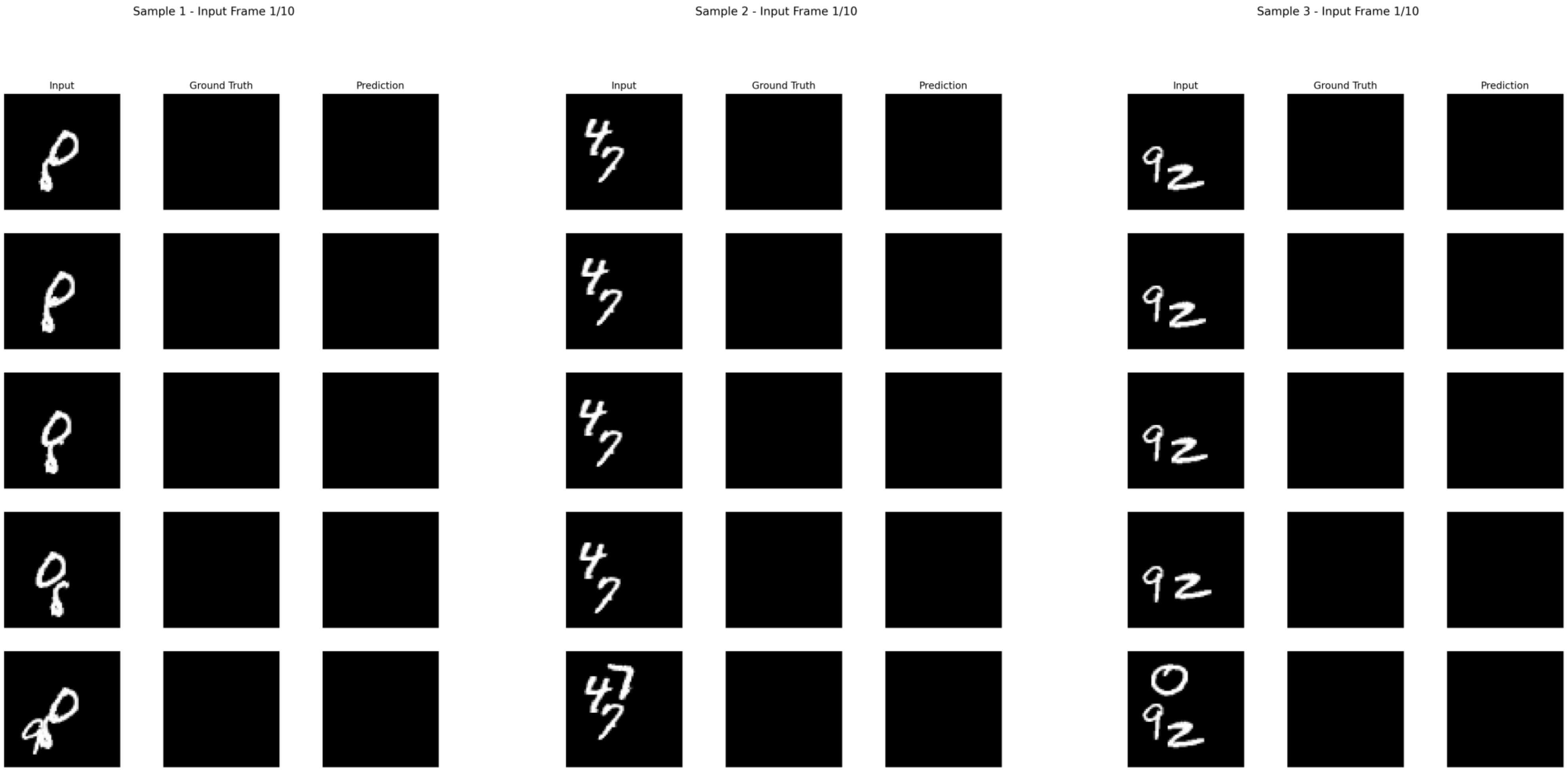
Original

1.5x speed

2x speed

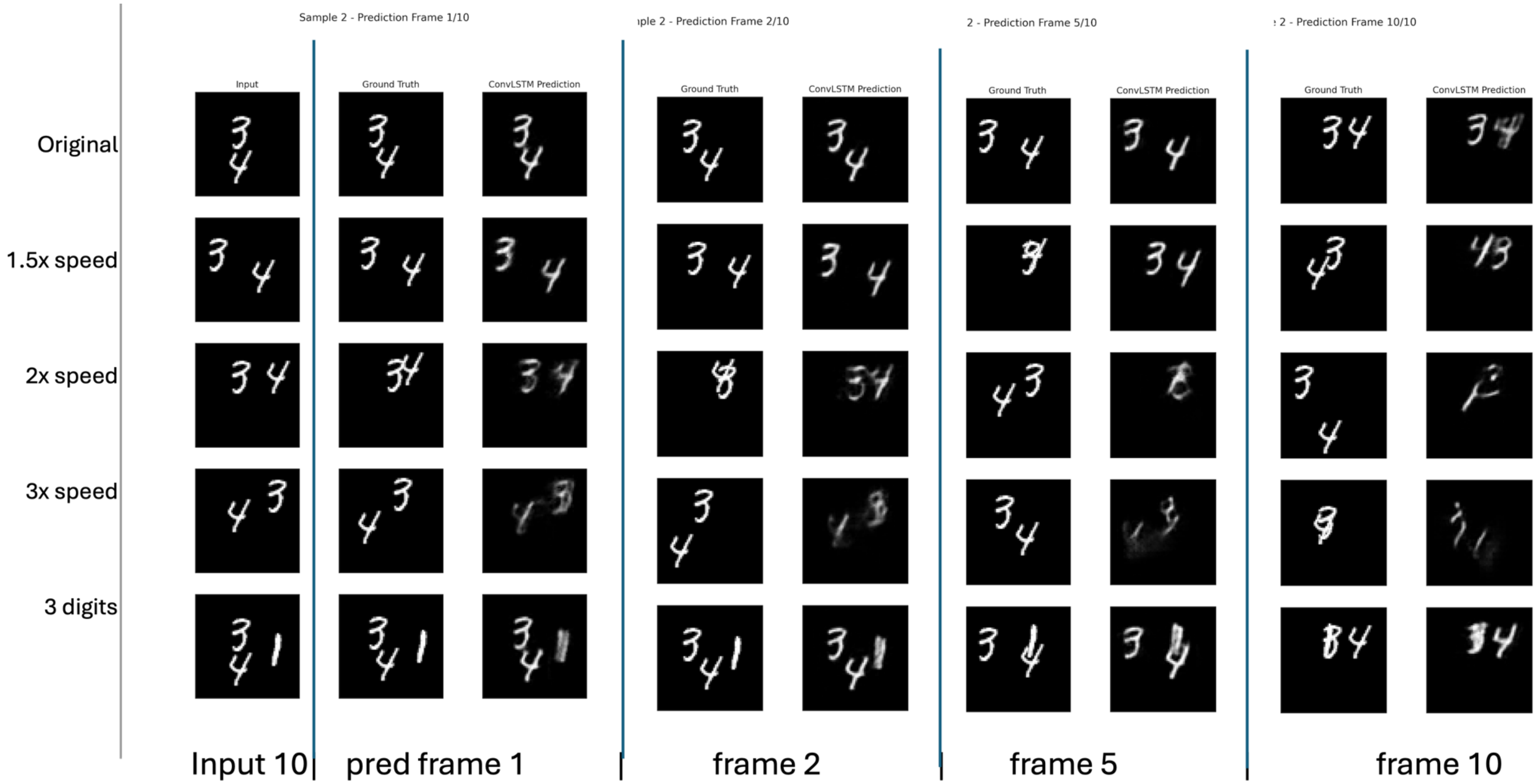
3x speed

3 digits


Sample 1 - Input Frame 1/10
Input
Ground Truth
Prediction
Sample 2 - Input Frame 1/10
Input
Ground Truth
Prediction
Sample 3 - Input Frame 1/10
Input
Ground Truth
Prediction

Zero-shot on Variants - convLSTM

Similar findings as SimVP,
but digits tend to disappear rather than become blurred..



Sample Movies from ConvLSTM

Original

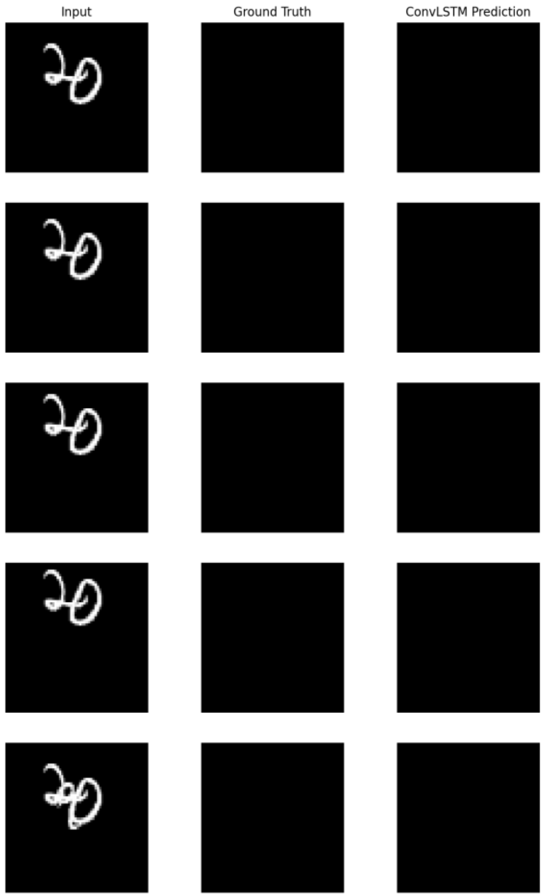
1.5x speed

2x speed

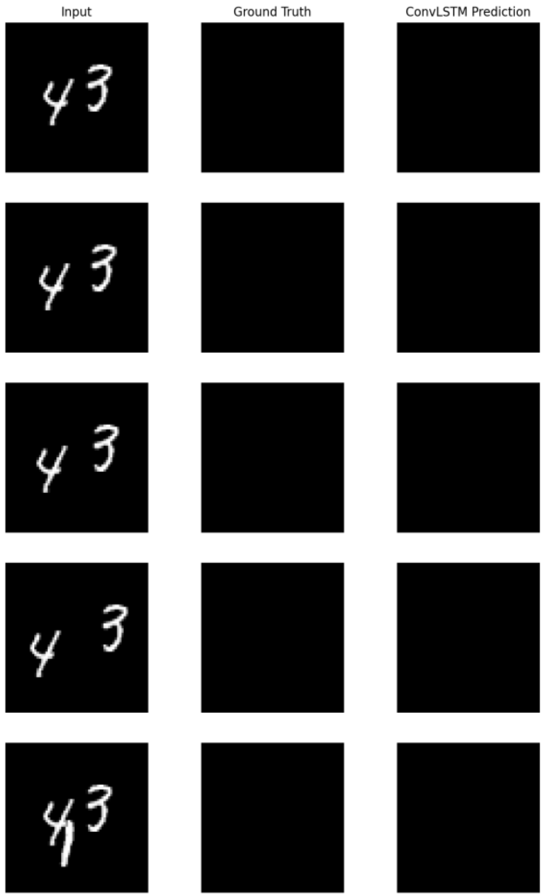
3x speed

3 digits

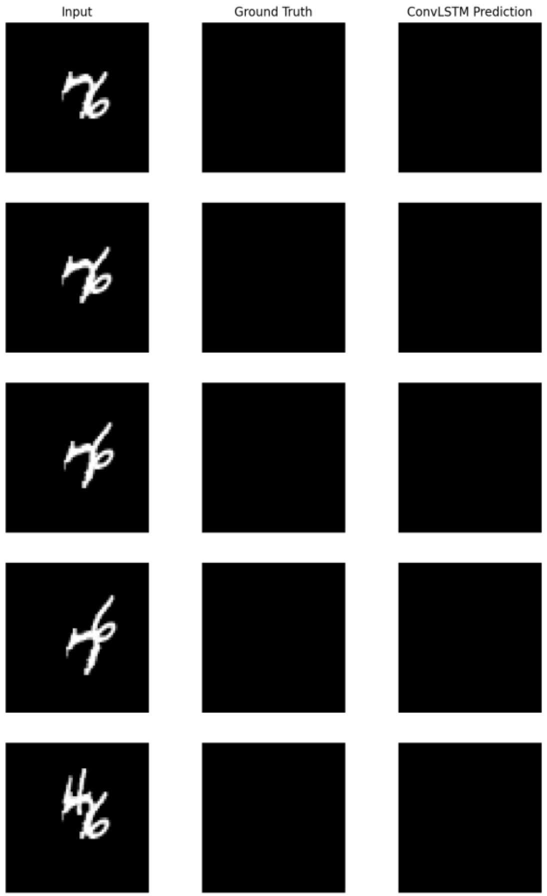
Sample 1 - Input Frame 1/10



Sample 2 - Input Frame 1/10



Sample 3 - Input Frame 1/10



Possible Next Step

- One shot or few shots to check the speed alignment.
- If the result improved, move to more stochastics motions like position-dependent sampling.