A APPENDIX

A.1 ABLATION STUDIES FOR TEMPORAL DYNAMICS ENCODER

We perform ablation studies on our model by trying different variants of recurrent modules for our temporal dynamics encoder networks. These models are: **DVG-RNN**, our model with an RNN dynamics encoder; **DVG-GRU**, with an RNN dynamics encoder with GRU units.

Figure A.1 shows ablation analysis for different variants of our approach. On the KTH dataset, different dynamics models (RNN, GRU, LSTM) all perform the same. On the BAIR dataset, RNN perform poorly and LSTM performs the best among the three. On Human3.6M dataset RNN performs higher than our LSTM and GRU models. On the FVD metric in Table A.1, all variants of our approach perform better than all baselines. In approaches, GRU dynamics model performs better on KTH and LSTM performs better on Human3.6M and BAIR dataset.

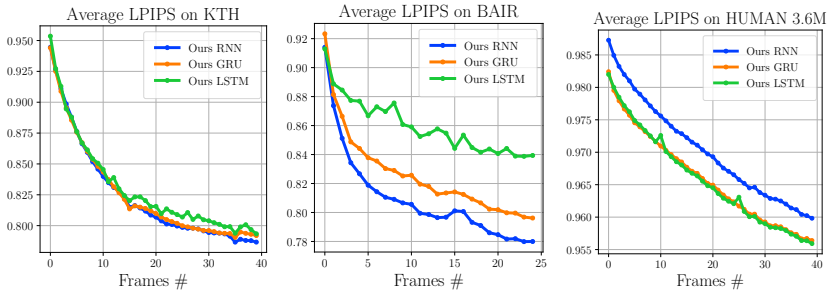

Figure A.1: **Ablation results** on KTH, Human3.6M and BAIR dataset using variants of temporal dynamics model in our method. We report best LPIPS metric. All methods use the best matching sample out of 100 random samples. We used fixed trigger to keep trigger point for each sample the same. On KTH, all temporal dynamics models have similar performance; and on BAIR, our LSTM model have best performance.

Table A.1: **Quantitative results** on KTH, BAIR, Human3.6M datasets. For the **FVD Score**, all the ablation methods use the best matching sample out of 100 random samples and lower numbers are better. For the **Diversity Score**, we compute the score across 50 generated samples, for 500 starting sequences, and higher numbers are better.

Model	Dynamics	Trigger	**FVD Score** (↓)			**Diversity Score** (↑) (frames: [10,25])		**Diversity Score** (↑) (frames: [25,40])	
			KTH	BAIR	Human3.6M	KTH	Human3.6M	KTH	Human3.6M
DVG [**ours**]	LSTM	@15,35	65.69	**123.08**	**479.43**	48.30	**9.3**	**46.20**	9.0
DVG [**ours**]	GRU	@15,35	**64.89**	124.38	485.96	**48.53**	8.5	44.23	**9.1**
DVG [**ours**]	RNN	@15,35	66.84	126.07	503.64	46.60	7.6	41.50	8.2

A.2 ANALYSIS: CHANGES IN ACTION AFTER GP TRIGGERING

KTH action dataset comprises of 6 action classes: walking, running, jogging, waving, clapping, and boxing. On an abstract level, we can cluster these actions into two categories moving actions and still actions. From the Figure A.2, it is interesting to observe that our GP triggering model captures the future trajectories of the videos and clusters them into these two categories moving actions and still actions. Common action switches that are to be expected can be observed from the Fig A.2; for example, walk and jog, wave and clap interchange frequently after triggering. Still actions seldom change to moving actions.

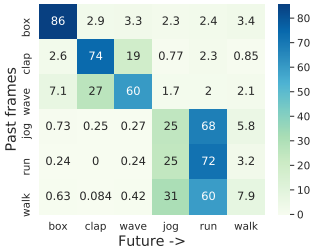

Figure A.2: **Changes in action from past frames to future frames** on KTH dataset. Total of 25,000 generated videos were used to calculate percentage change shown in the above figure.

A.3 SSIM AND PSNR RESULTS

We evaluated our generated video sequences using the tradition metrics like structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR) for comparison with previous baselines which reported these metrics. We trained all models on 64×64-size frames from the KTH, Human3.6M, and BAIR datasets. We used the standard training practice of using 5 frames as context (or past) and the model have to predict the next 10 frames. For all methods, SSIM and PSNR is computed by drawing 100 samples from the model for each test sequence and picking the best score with respect to the ground truth. We emphasize that these results are only for completeness and we hope that the community will stop relying on such reconstruction metrics for video prediction.

Results are reported in Figure A.3 represent the evaluation plots for traditional metrics on KTH, BAIR, and Human3.6M dataset. We follow the experimental setups from the baseline papers.

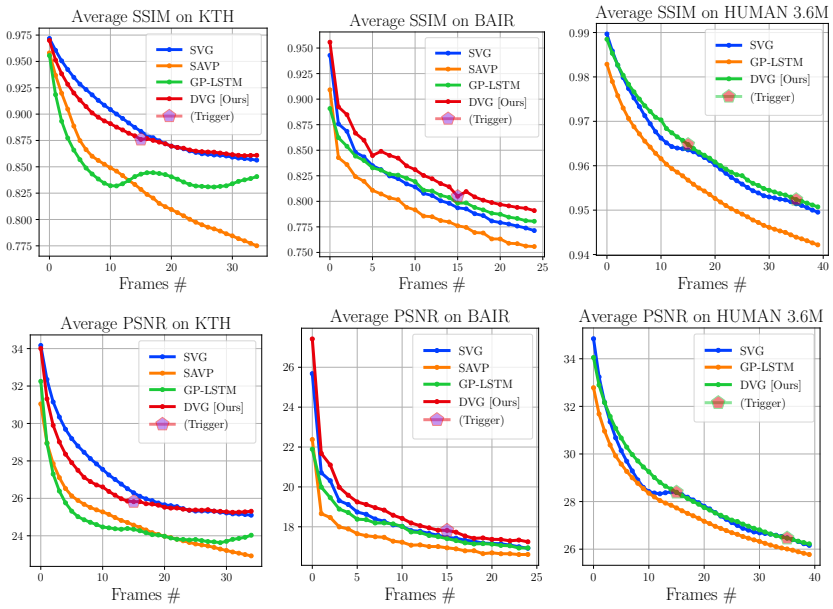

Figure A.3: **Quantitative results** on KTH, BAIR and Human3.6M dataset. We report average SSIM and PSNR metrics. All methods use the best matching sample out of 100 random samples. We used fixed trigger to keep trigger point for each sample the same.

A.4 QUALITATIVE RESULTS

It can be observed from Fig A.4 that after 15th frame SVG-LP is stuck in the same pose while after 35th frame SAVP starts distorting the human. However, our method (DVG) consistently generates frames that are diverse and distortion free for longer period of time. Similarly, in Fig A.9 it can be observed that after 30th frame SVG-LP and SAVP start generating subpar frames while our method is able to generate visually acceptable sequences for longer term. Few additional qualitative results on the BAIR dataset are provided in Figs A.5-A.6, and on the Human3.6M dataset in Figs A.7-A.8.

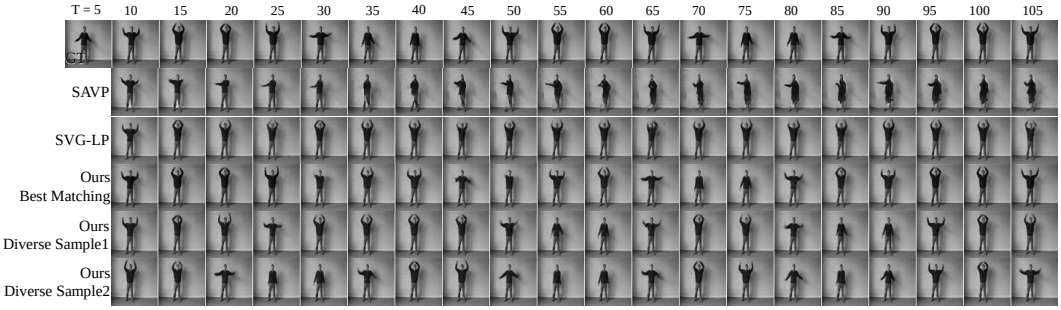

Figure A.4: **KTH dataset:** Qualitative comparison of the generated video sequences (every 5^{th} frame shown). First row is the ground-truth video (with last frame of the provided 5 frames is shown)

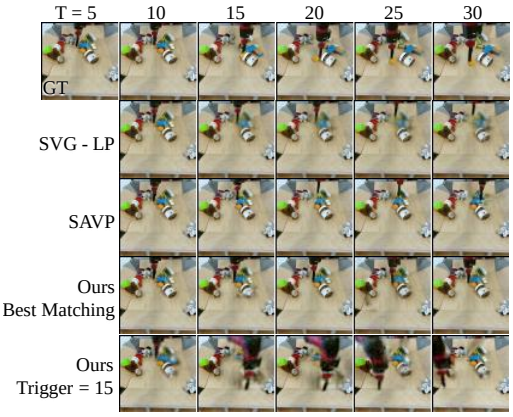

Figure A.5: **Qualitative results** on BAIR dataset. We show the best LPIPS samples out of 100 samples for all methods.

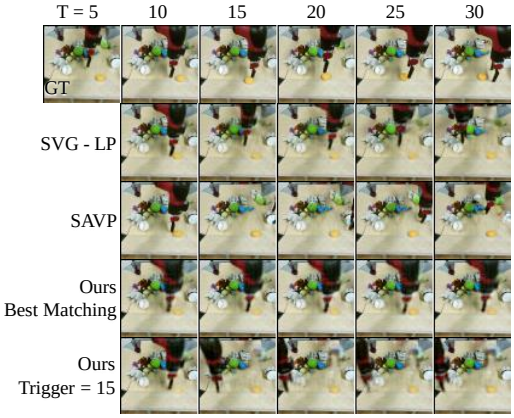

Figure A.6: **Qualitative results** on BAIR dataset. We show the best LPIPS samples out of 100 samples for all methods.

Figure A.7: **Human3.6M dataset:** Qualitative comparison of the generated video sequences (every 5^{th} frame shown). First row is the ground-truth video (with last frame of the provided 5 frames is shown)

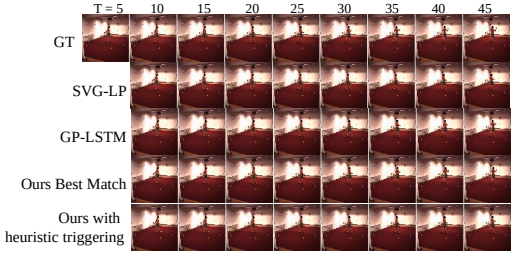

Figure A.8: **Human3.6M dataset:** Qualitative comparison of the generated video sequences (every 5^{th} frame shown). First row is the ground-truth video (with last frame of the provided 5 frames is shown)

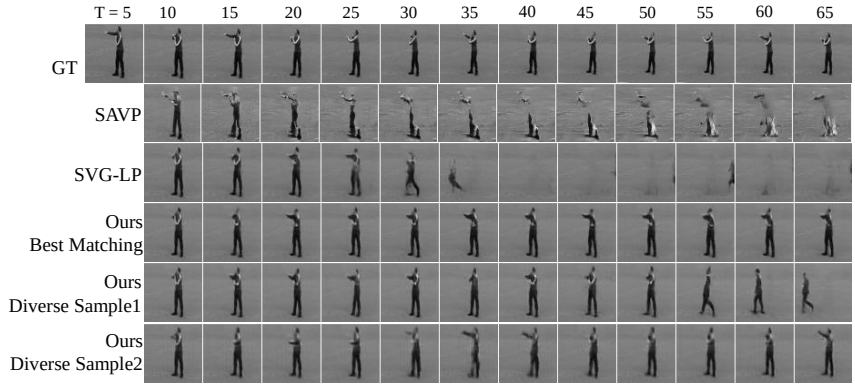

Figure A.9: **KTH dataset:** Qualitative comparison of the generated video sequences (every 5^{th} frame shown). First row is the ground-truth video (with last frame of the provided 5 frames is shown)

A.5 Gaussian Layer Specifics

As mentioned in the paper, GPytorch was used for our GP layer implementation. We utilized a large-scale variational GP implementation of GPytorch for our multi-dimensional GP regression problem of learning to predict the variance over the future frames in the latent space. For variational GP implementation, 40 inducing points were randomly initialized and learned during the training of GP. We used a RBF kernel along with gaussian likelihood for our GP layer. For optimization of our GPLayer, we employed stochastic optimization technique (Adam optimizer) to minimize the variational ELBO for a GP.

A.6 I3D Network architecture for Action Classifier

For our diversity metric mentioned in §5, we utilized the standard kinetics-pretrained I3D action recognition classifier. The input to the action classifier is a 15 frames clip and each frame has a size of 64×64. The action classifier attains accuracy close to 100% for KTH dataset and is above 90% accuracy for human3.6m dataset.