ICLR

Diverse Video Generation using a Gaussian Process Trigger

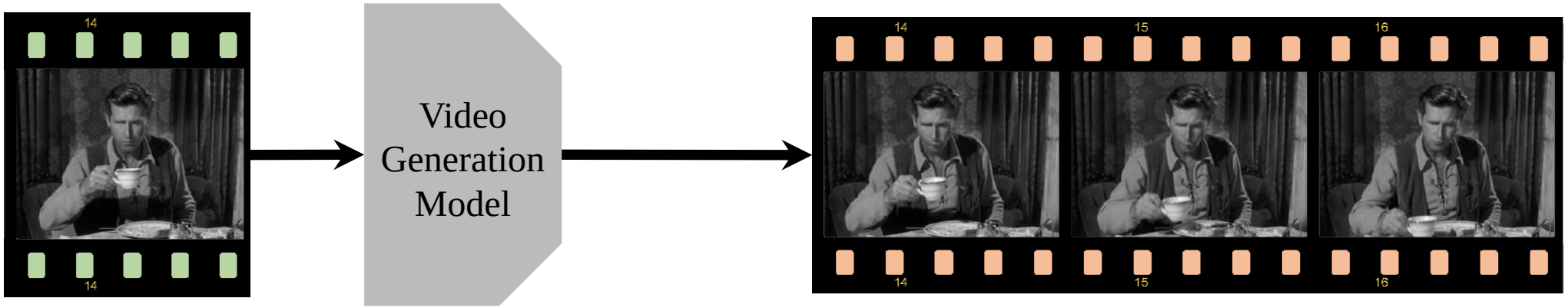
Gaurav Shrivastava and Abhinav Shrivastava

University of Maryland, College Park

http://www.cs.umd.edu/~gauravsh/dvg.html

Video Generation

Objective: Generate future frames given a few context (or past) frames.

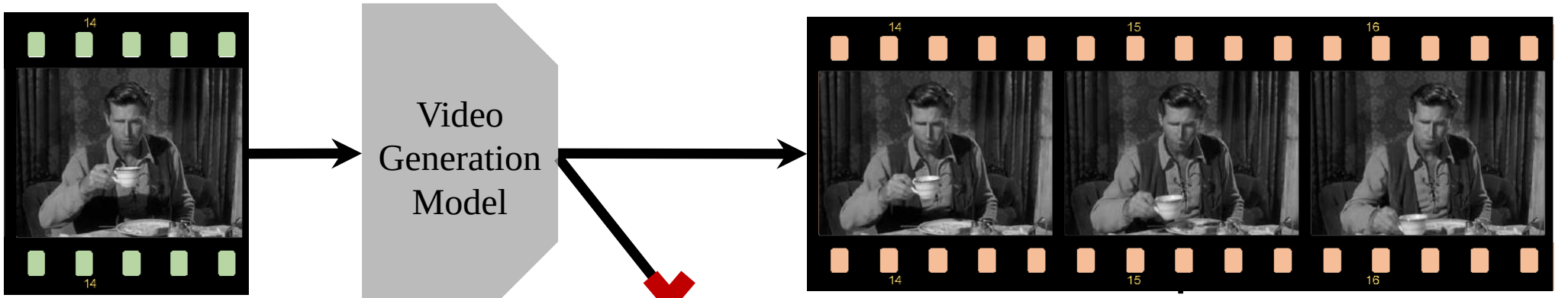


Past Frames
(Person Holding Cup)

Ground Truth
(Person Keeping the cup on the table)

Current Video Generation Methods

Focuses only on a single ground truth rather than producing a different plausible video sequence.



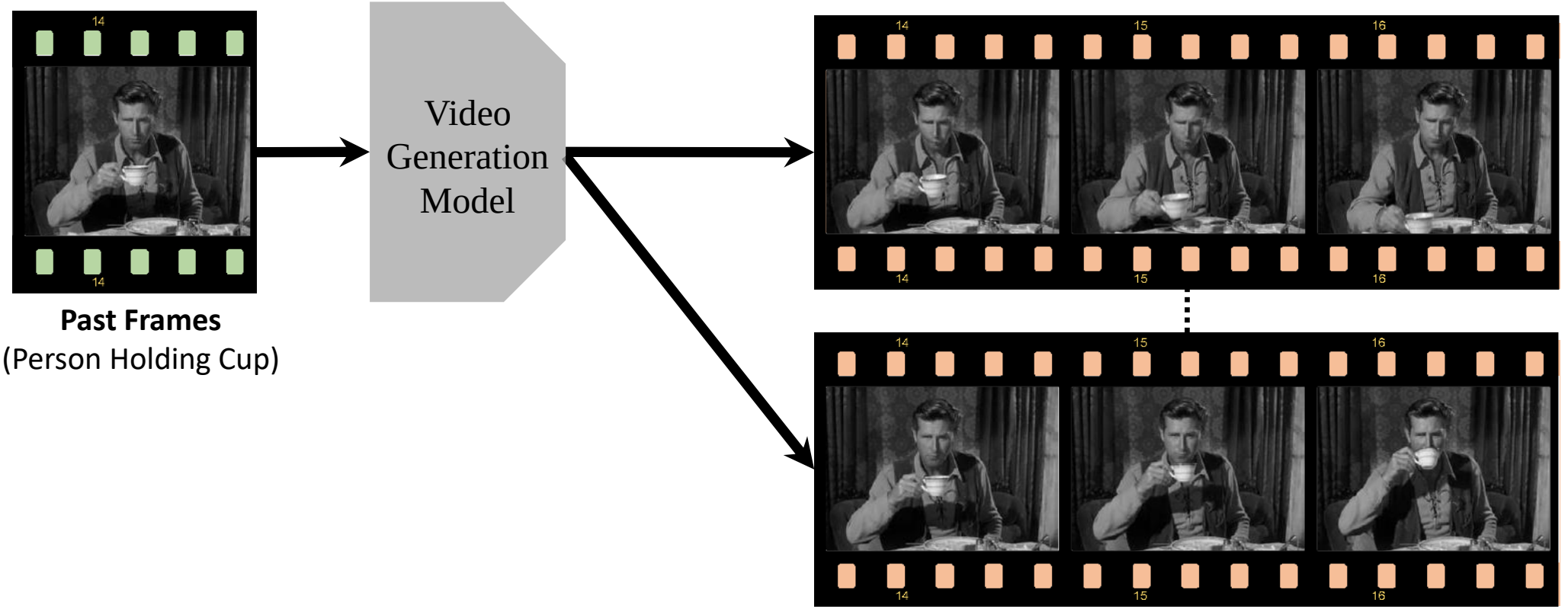
Past Frames
(Person Holding Cup)

Only Ground Truth
(Person Keeping the cup on the table)

Plausible Sample
(Person drinking from the cup)

Diverse Video Generation

Objective: Generate diverse future frames given a few context (or past) frames.



Background

Gaussian Process

A GP is a (potentially infinte) collection of random variables (RV) such that the joint distribution of every finite subset of RVs is multivariate Gaussian.

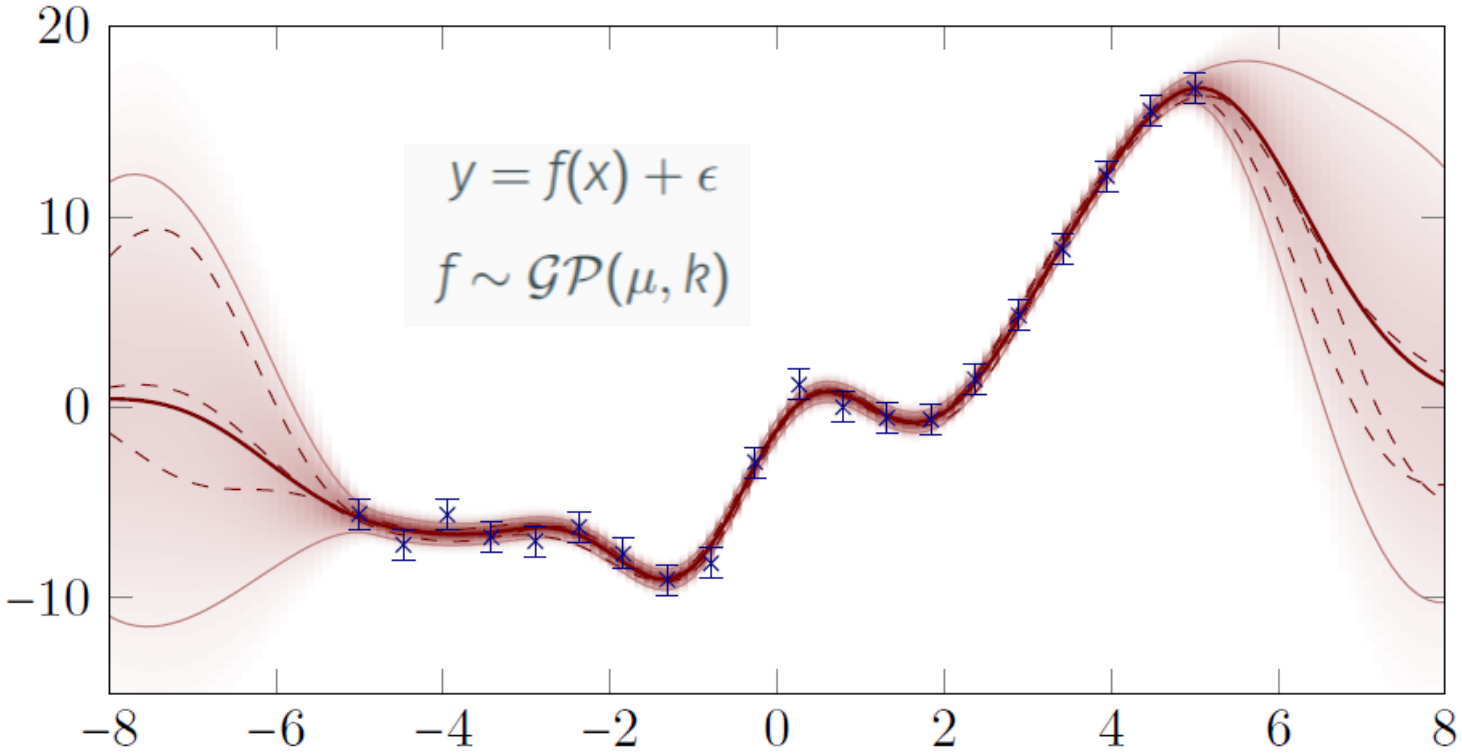
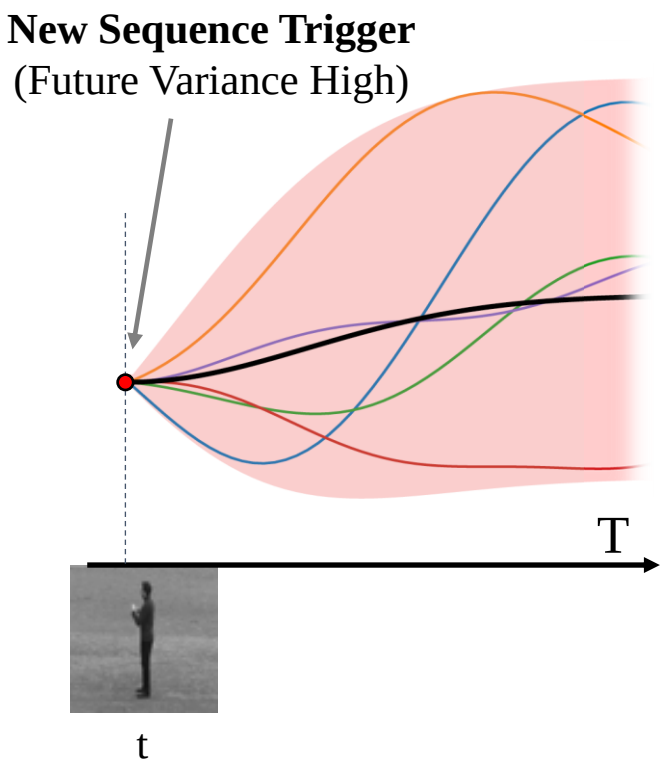
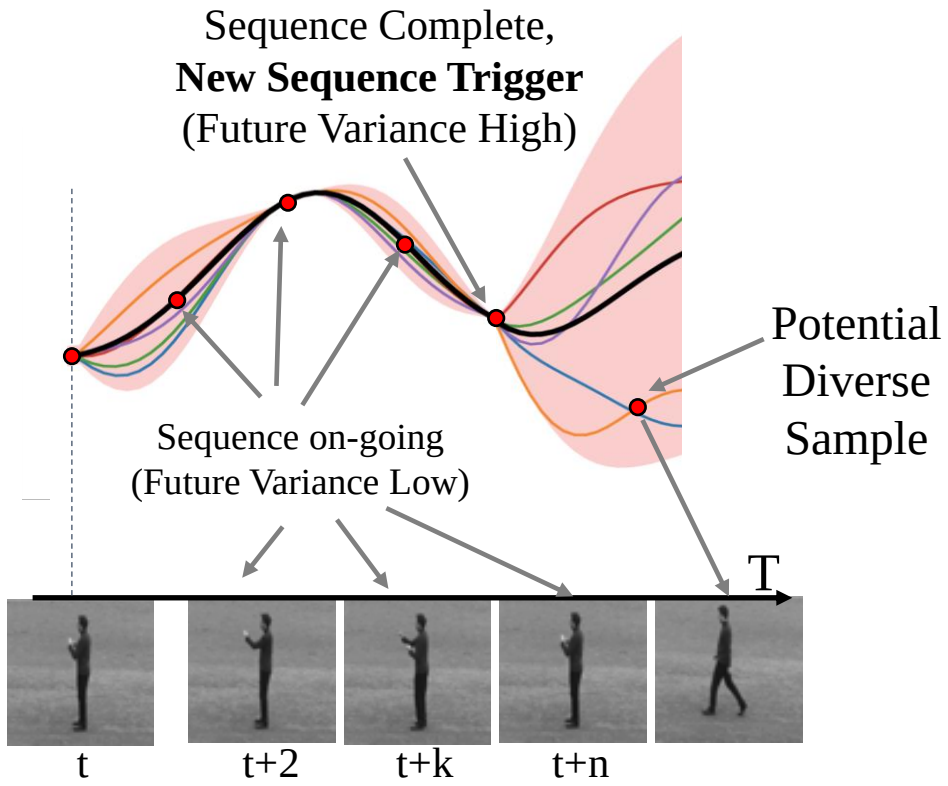


image source

Approach

Intuition: How to learn these diverse predictions from data?

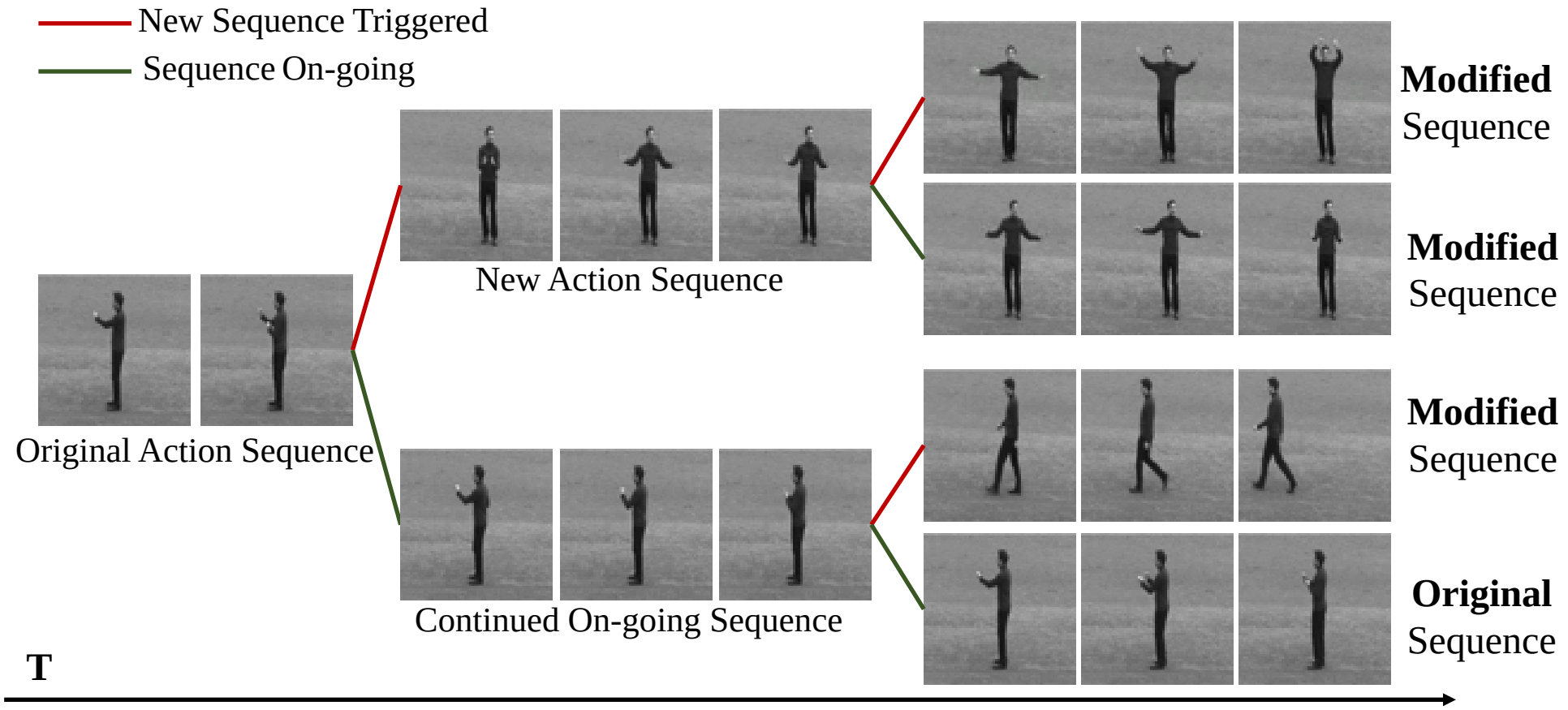




Key Insights:

- Predictive variance of GP is low when in between an action sequence.
- Predictive variance of GP is high when action sequence is complete.

Illustration of Our Approach



Components

Our model architecture consists of **three** main components:

Frame Auto-Encoder Network

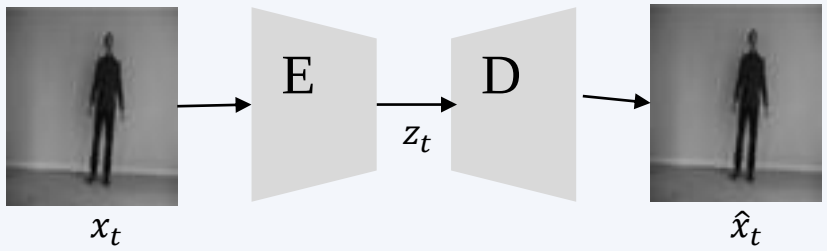
Maps input image($X_t \in R^{H \times W}$) to lower dimensional latent space($z_t \in R^D$) and back

LSTM Dynamics Encoder

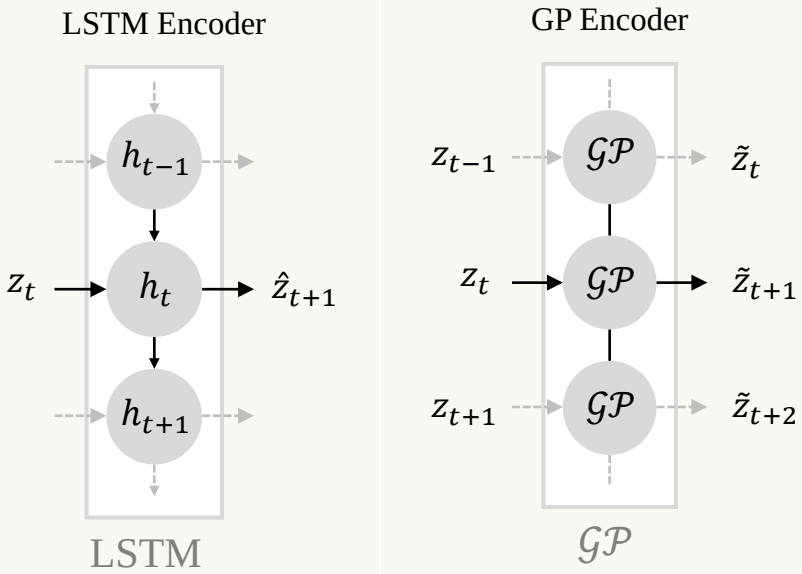
Encodes the dynamics of on-going actions in latent space

GP Dynamics Encoder

Maintains predictive variance over the future frames in latent space



Frame Auto-Encoder



Dynamics Encoders

Objective

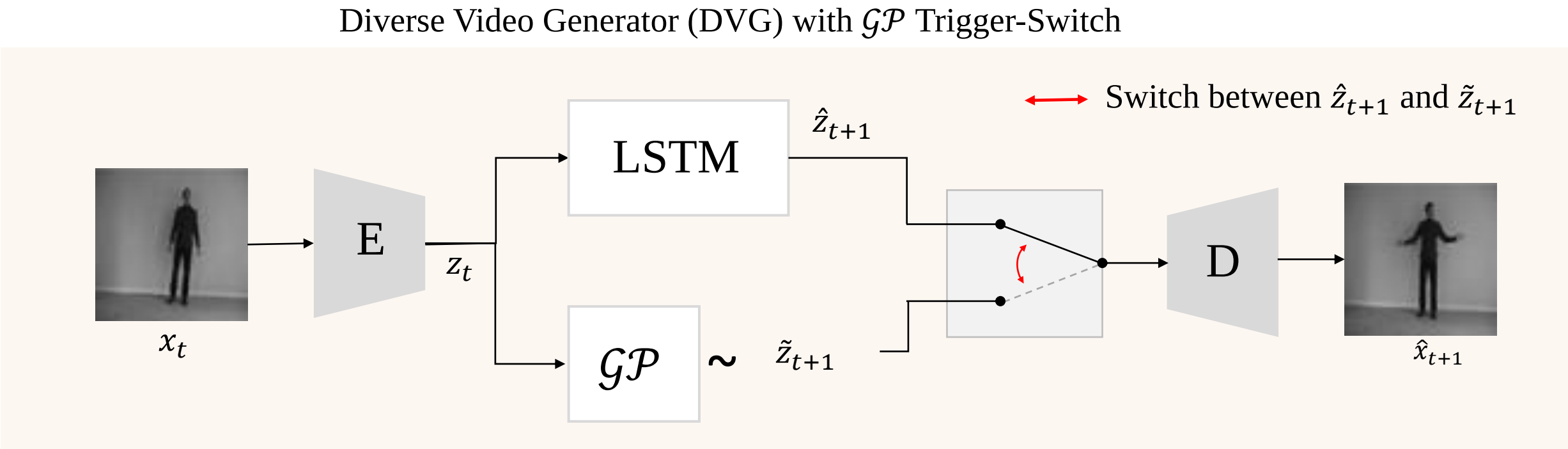
$$\mathcal{L}_{\text{DVG}} = \sum_{t=1}^{T} \Big(\overbrace{\lambda_1 \mathcal{L}_{\text{gen}}(\mathbf{x}_t, \hat{\mathbf{x}}_t)}^{\text{Frame Auto-Encoder}} + \overbrace{\lambda_2 \mathcal{L}_{\text{gen}}\left(\mathbf{x}_t, f_{\text{gen}}(\hat{\mathbf{z}}_t)\right)}^{\text{LSTM Frame Generation}} + \overbrace{\lambda_3 \mathcal{L}_{\text{gen}}\left(\mathbf{x}_t, f_{\text{gen}}(\tilde{\mathbf{z}}_t)\right)}^{\mathcal{GP}\text{ Frame Generation}} + \underbrace{\lambda_4 \mathcal{L}_{\text{LSTM}}(\mathbf{z}_{t+1}, \hat{\mathbf{z}}_{t+1})}_{\text{LSTM Dynamics Encoder}} + \underbrace{\lambda_5 \mathcal{L}_{\text{GP}}(\mathbf{z}_{t+1}, \mathbf{z}_t)}_{\mathcal{GP}\text{ Dynamics Encoder}} \Big)$$

where $[\lambda 1, \ldots, \lambda 5]$ are hyperparameters.

All the components are trained jointly.

- **Three frame generation losses**
  - z_t from frame encoder, $\hat{z}_t$ from LSTM encoder, and $\tilde{z}_t$ from GP encoder
- One LSTM dynamics encoder network loss
- One GP dynamics encoder network loss

Inference Architecture



Trigger Switch: The variance of GP is used to decide if we want to continue an on-going action or generate new diverse output.

Results

Evaluation of Generated Sequences

We evaluated our model against the state of the art models for video generation on the following

Quantitative metrics.

- **Accuracy of Reconstruction**
  - Traditional Metrics - *PSNR, SSIM*
  - Deep Metrics - *Perceptual Similarity (LPIPS), FVD*
- **Diversity of Sequences**
  - *Proposed new diversity metric*

Diversity Metric

Diversity Score: we compute mean number of generated clips that changed from the on-going action as classified by a trained classifier.

$$\text{Diversity Score} = \frac{1}{N}\sum_{N} I(c_i \neq \hat{c}_i)$$

Quantitative Results – KTH, BAIR, Human3.6M

Model	Trigger	**FVD Score** (↓)			**Diversity Score** (↑) (frames: [10,25])		**Diversity Score** (↑) (frames: [25,40])	
		KTH	BAIR	Human3.6M	KTH	Human3.6M	KTH	Human3.6M
SVG-LP	-	156.35	270.04	718.04	20.10	4.8	21.20	4.6
SAVP	-	65.98	126.75	-	26.60	-	24.50	-
GP-LSTM	-	92.34	197.49	604.75	31.40	5.4	30.90	6.0
VidFlow	-	-	124.81	-	. -	-	-	-
VRNN	-	67.26	134.81	523.45	32.50	-	31.80	-
DVG [**ours**]	@15,35	**65.69**	123.08	**479.43**	48.30	9.3	46.20	9.0
DVG [**ours**]	$\mathcal{GP}$	69.63	**120.03**	496.89	**47.71**	**10.8**	**48.10**	**10.1**

Table 1: **Quantitative results** on KTH, BAIR, Human3.6M datasets. For the **FVD Score**, all 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.

Qualitative Results

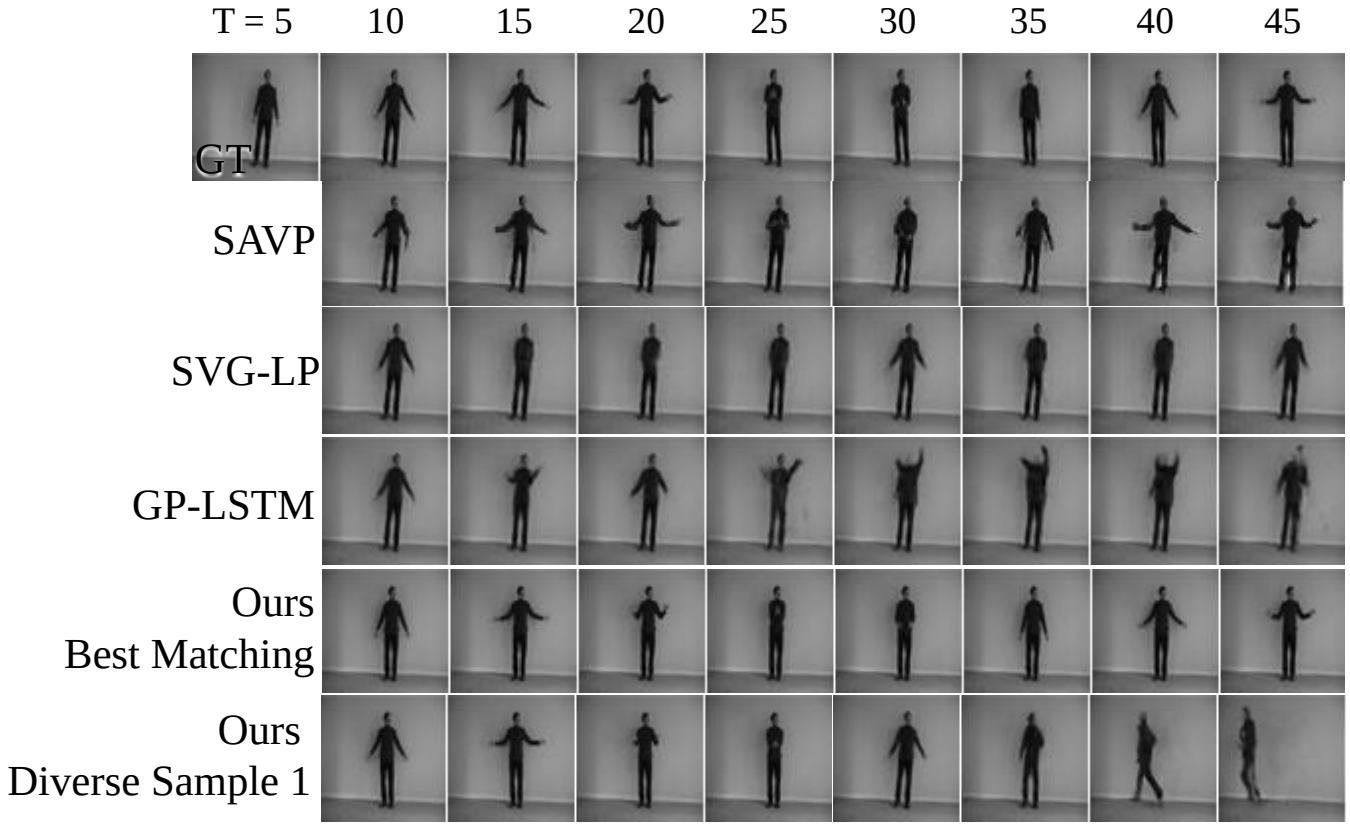
Every GP trigger tries to sample a diverse trajectory:

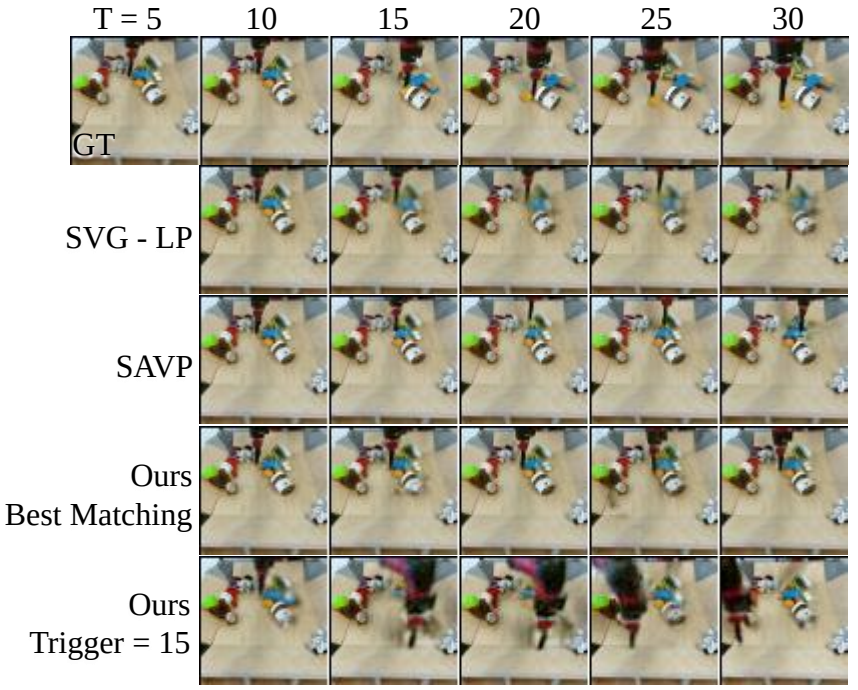
1) The person changes the course of direction at an angle
2) In this the person turns around at the trigger
3) The person tries to follow the ground truth
4) The person turns around on first trigger and on second trigger changes the action from walking to running

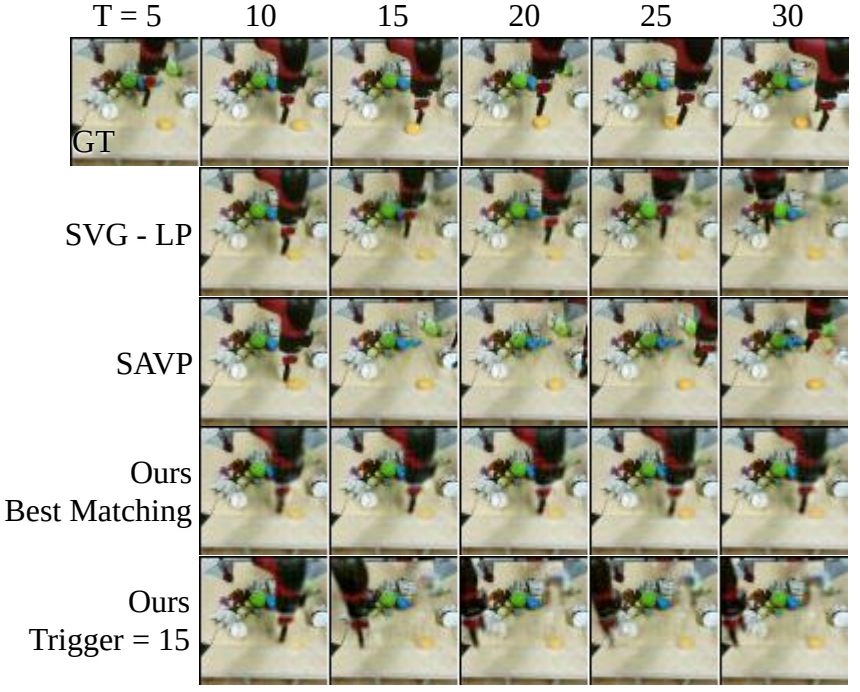
Interesting point to note: This turn around action is actually not present in the dataset and is the new action plausible that is derived by the model!

More Visualizations can be found on our webpage: http://www.cs.umd.edu/~gauravsh/dvg.html

Qualitative Results





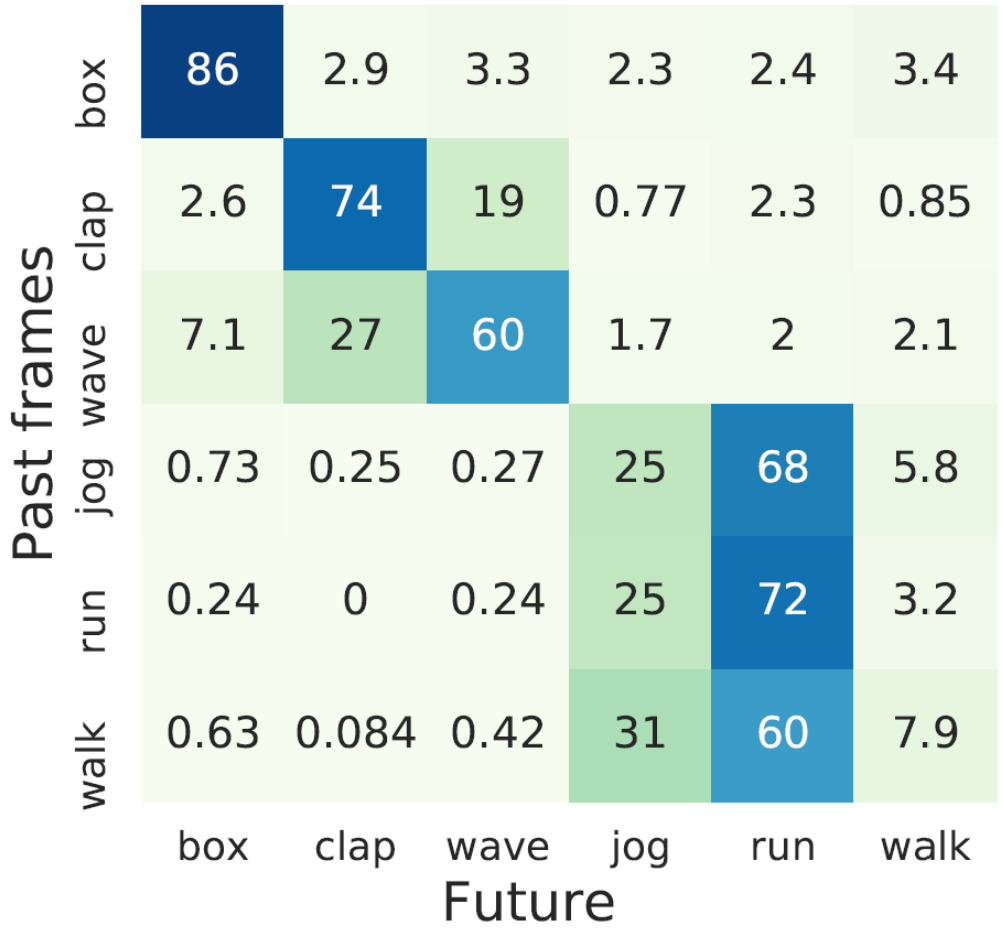


Action Transition Analysis

Two clusters in generated sequences:

- Moving actions (walk, run, jog)
- Standing actions (wave, clap, box)

Moving actions ✖ Standing actions



Questions?

- Please come visit us during our poster session!
- Poster Session: 4th May 8pm – 10pm EST

References:

Görtler, et al., "A Visual Exploration of Gaussian Processes", Distill, 2019.