Supplementary Material: Improved Conditional VRNNs for Video Prediction

A. Hierarchical VRNN

A.1. ELBO Derivation

We start from the ELBO for VRNNs [3]:

$$\log p(\mathbf{x}|\mathbf{c}) \geq \sum_{t=1}^{T} \mathbb{E}_{q(z_t|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})} \log p(x_t|\mathbf{z_{\leq t}},\mathbf{x_{<t}},\mathbf{c}) - D_{KL}(q(z_t|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})||p(z_t|\mathbf{z_{<t}},\mathbf{x_{<t}},\mathbf{c})) \tag{6}$$

Recall we defined $\mathbf{z_t} = (z_t^1, ..., z_t^L)$ and factorized the prior as:

$$p(\mathbf{z_t}|\mathbf{z_{<t}},\mathbf{x_{<t}},\mathbf{c}) = \prod_{l=1}^{L} p(z_t^l|\mathbf{z_t^{<l}},\mathbf{z_{<t}^l},\mathbf{x_{<t}},\mathbf{c}). \tag{7}$$

And the posterior:

$$q(\mathbf{z_t}|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c}) = \prod_{l=1}^{L} q(z_t^l|\mathbf{z_t^{<l}},\mathbf{z_{<t}^l},\mathbf{x_{\leq t}},\mathbf{c}). \tag{8}$$

We then substitute these terms in the VRNN ELBO, first looking at the reconstruction term inside the summation over time:

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{z_t}|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})} \log p(x_t|\mathbf{z_{\leq t}},\mathbf{x_{<t}},\mathbf{c}) &= \sum_{t=1}^{T} \mathbb{E}_{q(z_t^1,...,z_t^L|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})} \log p(x_t|z_t^1,...,z_t^L,\mathbf{z_{<t}},\mathbf{x_{<t}},\mathbf{c}) \\ &= \sum_{t=1}^{T} \mathbb{E}_{q(z_t^1|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})...q(z_t^L|\mathbf{z_t^{<L}},\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})} \log p(x_t|z_t^1,...,z_t^L,\mathbf{z_{<t}},\mathbf{x_{<t}},\mathbf{c})\end{aligned} \tag{9}$$

And then looking at the summation of KL divergences:

$$\begin{aligned}-\sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{z_t}|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})} \log \frac{q(\mathbf{z_t}|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})}{p(\mathbf{z_t}|\mathbf{z_{<t}},\mathbf{x_{<t}},\mathbf{c})} &= -\sum_{t=1}^{T} \mathbb{E}_{q(z_t^1,...,z_t^L|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})} \log \frac{q(z_t^1,...,z_t^L|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})}{p(z_t^1,...,z_t^L|\mathbf{z_{<t}},\mathbf{x_{<t}},\mathbf{c})} \\ &= -\sum_{t=1}^{T} \mathbb{E}_{q(z_t^1|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})...q(z_t^L|\mathbf{z_t^{<L}},\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})} \log \frac{q(z_t^1|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})...q(z_t^L|\mathbf{z_t^{<L}},\mathbf{z_{<t}},\mathbf{x_{\leq t}})}{p(z_t^1|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})...q(z_t^L|\mathbf{z_t^{<L}},\mathbf{z_{<t}},\mathbf{x_{<t}})} \\ &= -\sum_{t=1}^{T} \mathbb{E}_{q(z_t^1|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})...q(z_t^L|\mathbf{z_t^{<L}},\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})} \log \frac{q(z_t^1|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})...q(z_t^L|\mathbf{z_t^{<L}},\mathbf{z_{<t}},\mathbf{x_{\leq t}})}{p(z_t^1|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})...q(z_t^L|\mathbf{z_t^{<L}},\mathbf{z_{<t}},\mathbf{x_{<t}})} \\ &\text{(by definition of conditional KL divergence)} \\ &= -\sum_{t=1}^{T}\sum_{l=1}^{L} D_{KL}((q(z_t|\mathbf{z_{<t}},\mathbf{x_{\leq t}},\mathbf{c})||p(z_t|\mathbf{z_{<t}},\mathbf{x_{<t}},\mathbf{c}))\end{aligned} \tag{10}$$

Adding both terms together we obtain the ELBO defined in eq. 5.

A.2. Posterior Dense Connectivity

Fig A.1 illustrates the dense connection of the approximate posterior. For each latent variable has a deterministic connection to x_{t-1} (red arrows in Fig 2), in addition to all the latent variables from the layers below (green arrow in in Fig 2). Finally , each latent variable has a direct connection to the output variables x_t, corresponding to the inference path.

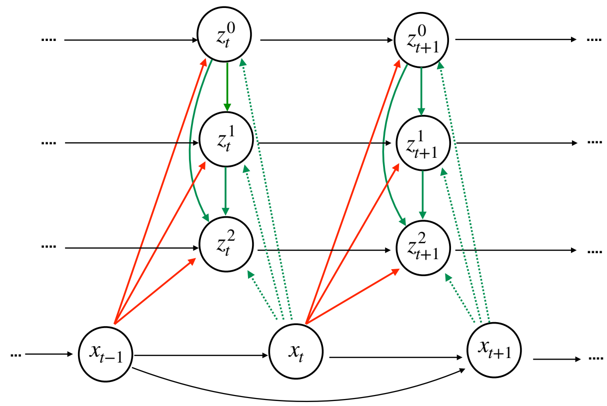


Figure A.1: **Schematic view of the approximate posterior with the dense-connectivity pattern**. Arrows in red show the connections from the input at the previous timestep to current latent variables. Arrows in green highlight skip connections between latent variables to outputs. Arrows in **black** indicate recurrent temporal connections. We empirically observe that this dense-connectivity pattern eases the training of latent hierarchy.

LAYERS	DETAILS
Conv2D	input $\rightarrow$ 64
ResNet Block	64 $\rightarrow$ 64
MaxPool	2x2, $s = 2$
ResNet Block	64 $\rightarrow$ 128
ResNet Block	128 $\rightarrow$ 128
MaxPool	2x2, $s = 2$
ResNet Block	128 $\rightarrow$ 256
ResNet Block	256 $\rightarrow$ 256
MaxPool	2x2, $s = 2$
ResNet Block	256 $\rightarrow$ 512
ResNet Block	512 $\rightarrow$ 512
MaxPool	2x2, $s = 2$
ResNet Block	512 $\rightarrow$ 512, $ks = 4$
ResNet Block	512 $\rightarrow$ 512

Figure B.1: **Frame Encoder architecture.**

B. Model Specification

We specify the architecture used for the 64x64 model. Convolutional layers in our model use 3x3 kernels with stride s = 1 and padding p = 1 unless otherwise specified. We use modified Resnet blocks made up of two groups of ReLU + Conv2D + GroupNorm. GroupNorm layers use g = 16 groups. Transposed Convolutions use 4x4 kernels with stride s = 2 and padding p = 1, which upscales 2x the input tensor. ConvLSTM layers use 3x3 kernels with stride s = 1 and padding p = 1 and GroupNorm.

C. Additional Samples

See the figures below.

D. PredNet

We additionally compare to PredNet on Cityscapes using the official implementation. Note that PredNet is deterministic and can't model future uncertainty. The model is only able to correctly predict a few timesteps before becoming blurry, as the uncertainty increases with time. PredNet obtained a FVD score of 1079.19 and a LPIPS score of 0.397, while ours with the hierarchical model are 567.51 and 0.264 respectively (lower is better).

Layers	Details
ConvLSTM	$512 \rightarrow 512$, $ks = 4$
UpConv	$512 \rightarrow 512$, scale $= 4$
ConvLSTM	$512 \rightarrow 512$
UpConv	$512 \rightarrow 512$
ConvLSTM	$512 \rightarrow 512$
UpConv	$512 \rightarrow 256$
ConvLSTM	$256 \rightarrow 256$
UpConv	$256 \rightarrow 128$
ConvLSTM	$128 \rightarrow 128$
UpConv	$128 \rightarrow 64$
ConvLSTM	$64 \rightarrow 64$
Conv2D + GroupNorm + ReLU	$64 \rightarrow 64$
Conv2D	$64 \rightarrow$ input

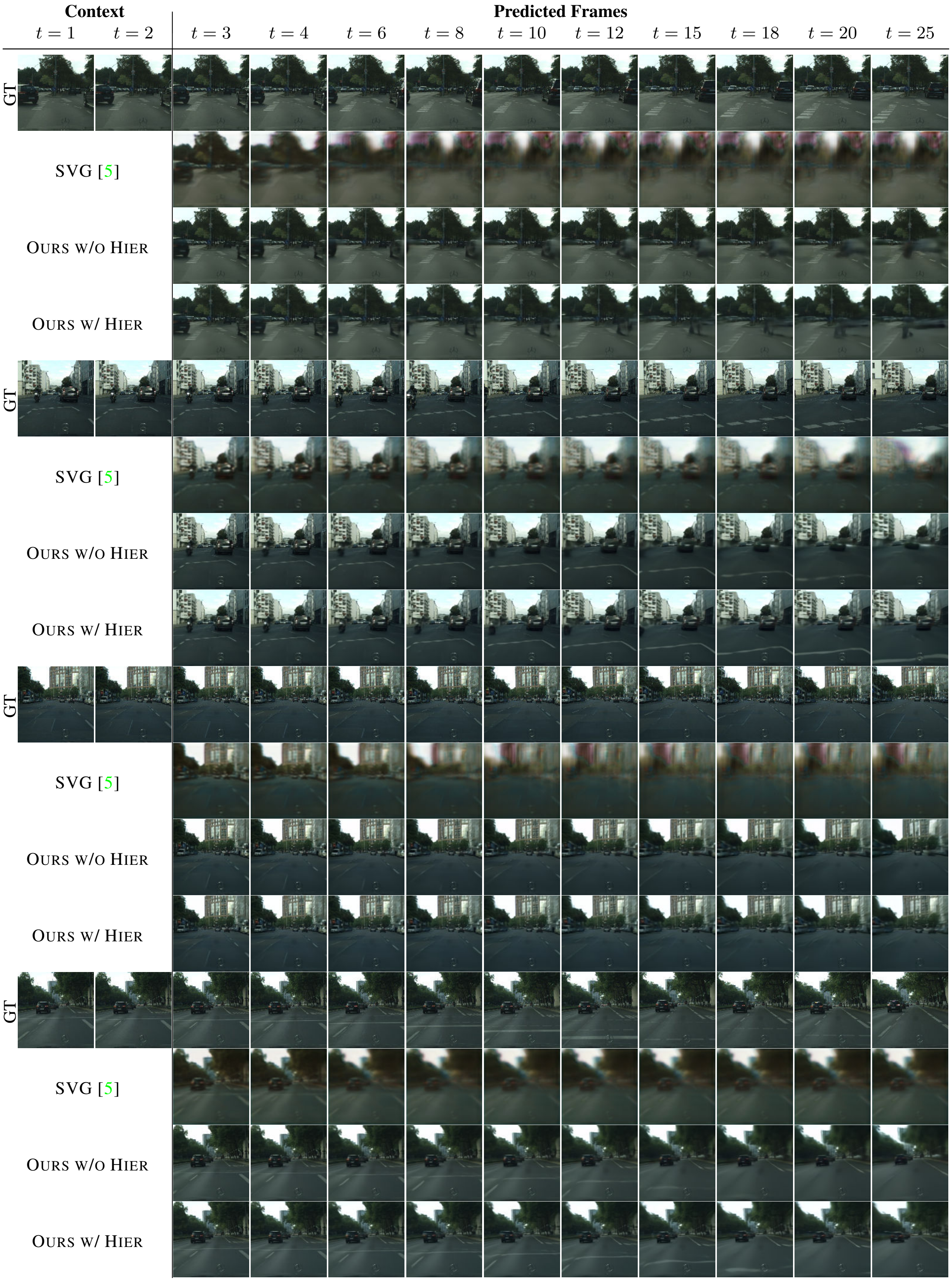
Figure B.2: **LIkelihood/Decoder architecture.**

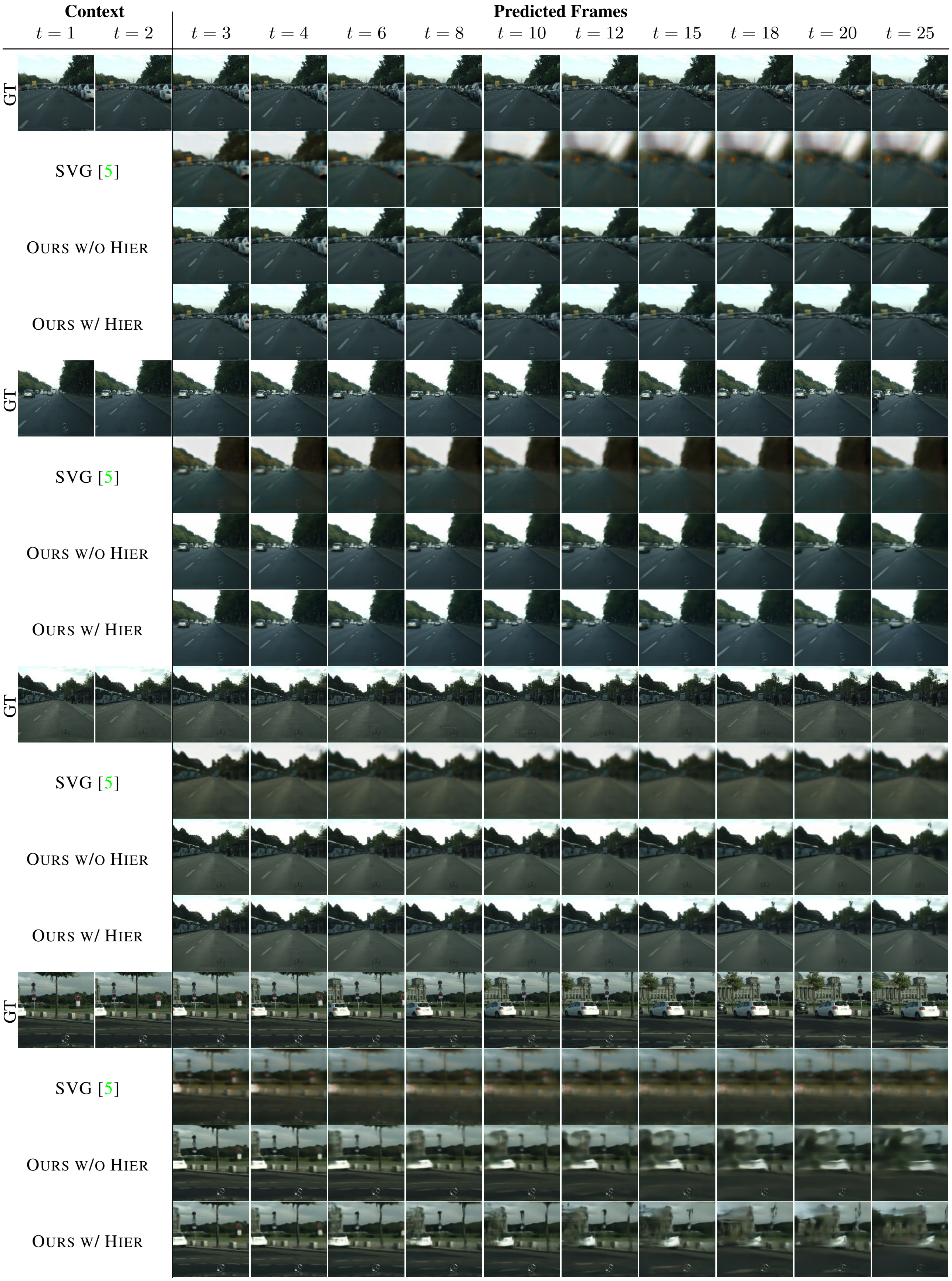
Level	Layers	Details
1x1	Conv2D + GroupNorm	$128 \rightarrow 128$, $ks = 1$
	ConvLSTM	$128 \rightarrow 128$
	Conv2D + GroupNorm	$128 \rightarrow 128$x2, $ks = 1$
8x8	Conv2D + GroupNorm	$512 \rightarrow 512$, $ks = 1$
	ConvLSTM	$512 \rightarrow 1512$
	Conv2D + GroupNorm	$512 \rightarrow 512$x2, $ks = 1$
32x32	Conv2D + GroupNorm	$512 \rightarrow 512$, $ks = 1$
	ConvLSTM	$512 \rightarrow 1512$
	Conv2D + GroupNorm	$512 \rightarrow 512$x2, $ks = 1$

Figure B.3: **Prior/Posterior architecture.**

Level	Layers	Details
1x1	Conv2D + GroupNorm + ReLU	$512 \rightarrow 512$, $ks = 1$
	Conv2D + GroupNorm	$512 \rightarrow 512$x2, $ks = 1$
4x4	Conv2D + GroupNorm + ReLU	$512 \rightarrow 512$, $ks = 1$
	Conv2D + GroupNorm	$512 \rightarrow 512$x2, $ks = 1$
8x8	Conv2D + GroupNorm + ReLU	$512 \rightarrow 512$, $ks = 1$
	Conv2D + GroupNorm	$512 \rightarrow 512$x2, $ks = 1$
16x16	Conv2D + GroupNorm + ReLU	$256 \rightarrow 256$, $ks = 1$
	Conv2D + GroupNorm	$256 \rightarrow 256$x2, $ks = 1$
32x32	Conv2D + GroupNorm + ReLU	$128 \rightarrow 128$, $ks = 1$
	Conv2D + GroupNorm	$128 \rightarrow 128$x2, $ks = 1$
64x64	Conv2D + GroupNorm + ReLU	$64 \rightarrow 64$, $ks = 1$
	Conv2D + GroupNorm	$64 \rightarrow 64$x2, $ks = 1$

Figure B.4: **Initial State network architecture.**


Context
Predicted Frames
$t=1$ $t=2$ $t=3$ $t=4$ $t=6$ $t=8$ $t=10$ $t=12$ $t=15$ $t=18$ $t=20$ $t=25$
GT
SVG [5]
OURS W/O HIER
OURS W/ HIER
GT
SVG [5]
OURS W/O HIER
OURS W/ HIER
GT
SVG [5]
OURS W/O HIER
OURS W/ HIER
GT
SVG [5]
OURS W/O HIER
OURS W/ HIER


Context
Predicted Frames
$t = 1$
$t = 2$
$t = 3$
$t = 4$
$t = 6$
$t = 8$
$t = 10$
$t = 12$
$t = 15$
$t = 18$
$t = 20$
$t = 25$
GT
SVG [5]
OURS W/O HIER
OURS W/ HIER
GT
SVG [5]
OURS W/O HIER
OURS W/ HIER
GT
SVG [5]
OURS W/O HIER
OURS W/ HIER
GT
SVG [5]
OURS W/O HIER
OURS W/ HIER

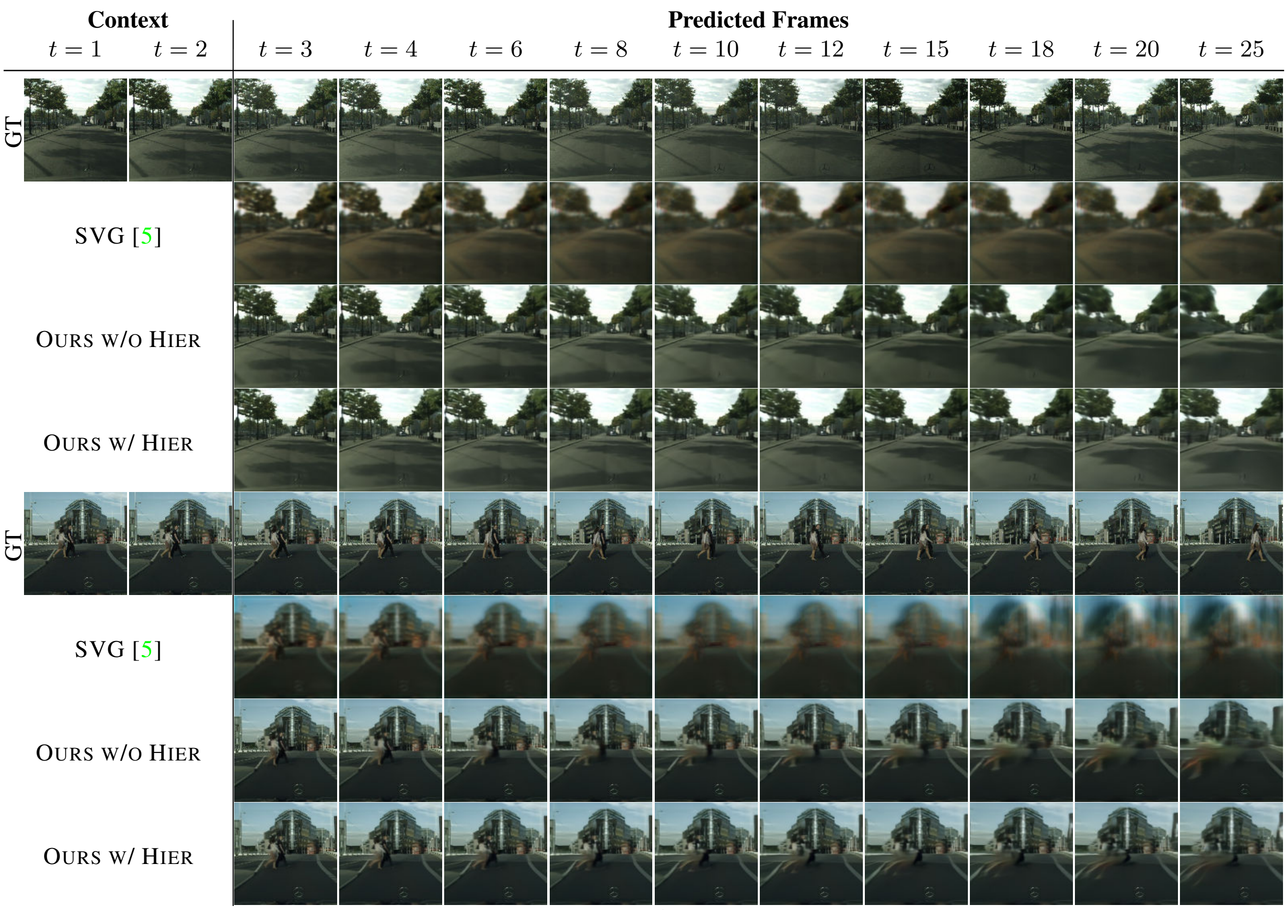
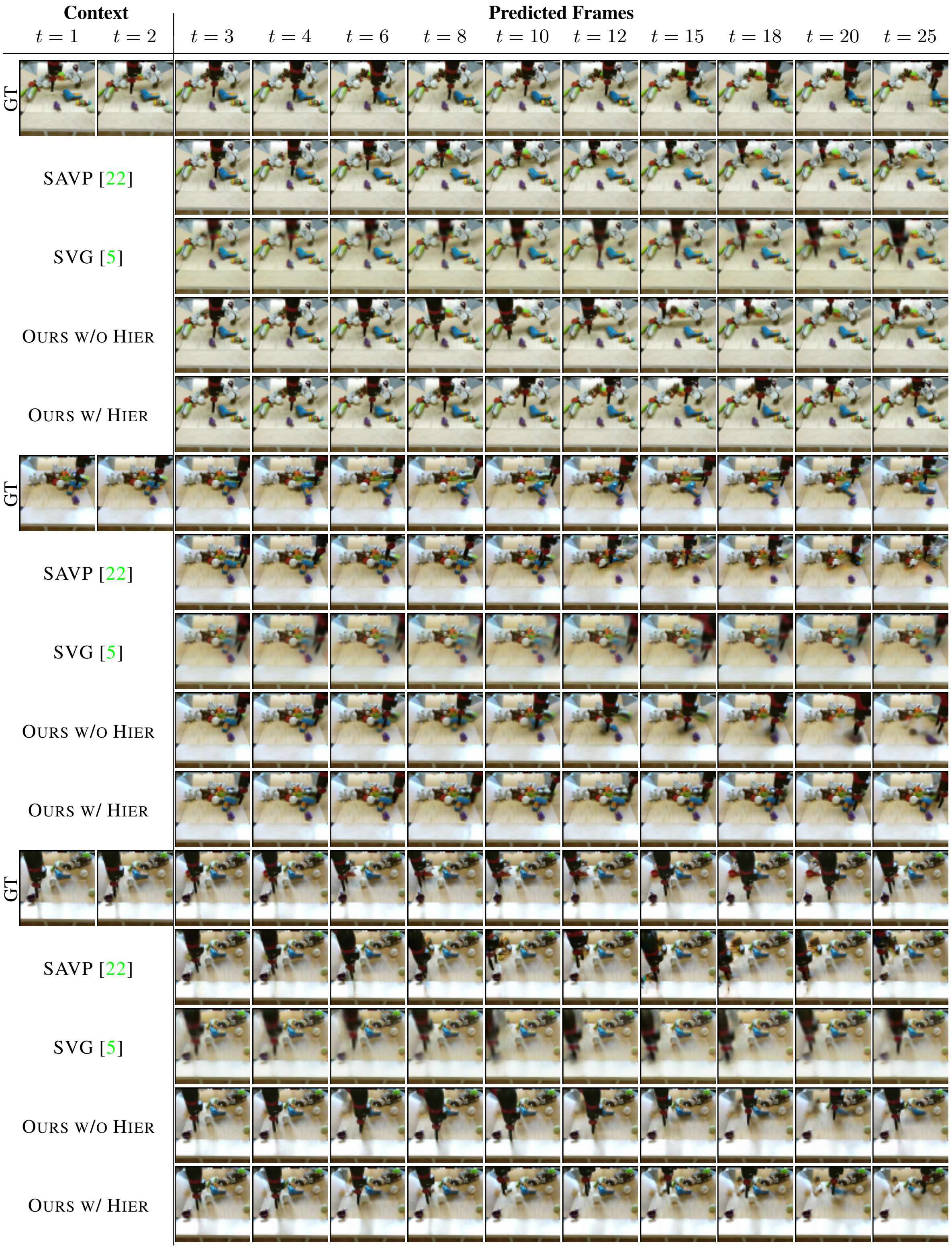


Figure C.1: **Random Test Samples for Cityscapes.**


Context
Predicted Frames
$t = 1$
$t = 2$
$t = 3$
$t = 4$
$t = 6$
$t = 8$
$t = 10$
$t = 12$
$t = 15$
$t = 18$
$t = 20$
$t = 25$
GT
SAVP [22]
SVG [5]
OURS W/O HIER
OURS W/ HIER
GT
SAVP [22]
SVG [5]
OURS W/O HIER
OURS W/ HIER
GT
SAVP [22]
SVG [5]
OURS W/O HIER
OURS W/ HIER

Context
Predicted Frames
$t = 1$
$t = 2$
$t = 3$
$t = 4$
$t = 6$
$t = 8$
$t = 10$
$t = 12$
$t = 15$
$t = 18$
$t = 20$
$t = 25$
GT
SAVP [22]
SVG [5]
OURS W/O HIER
OURS W/ HIER
GT
SAVP [22]
SVG [5]
OURS W/O HIER
OURS W/ HIER
GT
SAVP [22]
SVG [5]
OURS W/O HIER
OURS W/ HIER

Context
Predicted Frames
$t = 1$ $t = 2$ $t = 3$ $t = 4$ $t = 6$ $t = 8$ $t = 10$ $t = 12$ $t = 15$ $t = 18$ $t = 20$ $t = 25$
GT
SAVP [22]
SVG [5]
OURS W/O HIER
OURS W/ HIER
GT
SAVP [22]
SVG [5]
OURS W/O HIER
OURS W/ HIER
GT
SAVP [22]
SVG [5]
OURS W/O HIER
OURS W/ HIER

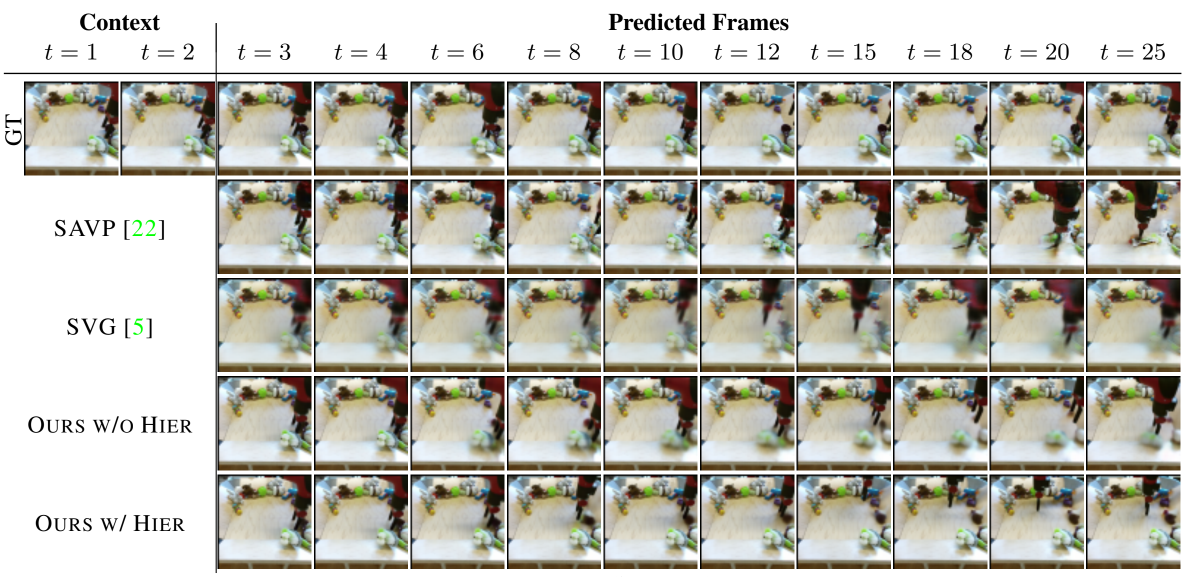


Figure C.2: **Random Test Samples for pushbair.**


Context
Predicted Frames
t = 1
t = 2
t = 3
t = 4
t = 6
t = 8
t = 10
t = 12
t = 15
t = 18
t = 20
GT
SVG [5]
OURS W/O HIER
OURS W/ HIER

Context
Predicted Frames
$t = 1$
$t = 2$
$t = 3$
$t = 4$
$t = 6$
$t = 8$
$t = 10$
$t = 12$
$t = 15$
$t = 18$
$t = 20$
GT
SVG [5]
OURS W/O HIER
OURS W/ HIER
GT
SVG [5]
OURS W/O HIER
OURS W/ HIER
GT
SVG [5]
OURS W/O HIER
OURS W/ HIER
GT
SVG [5]
OURS W/O HIER
OURS W/ HIER

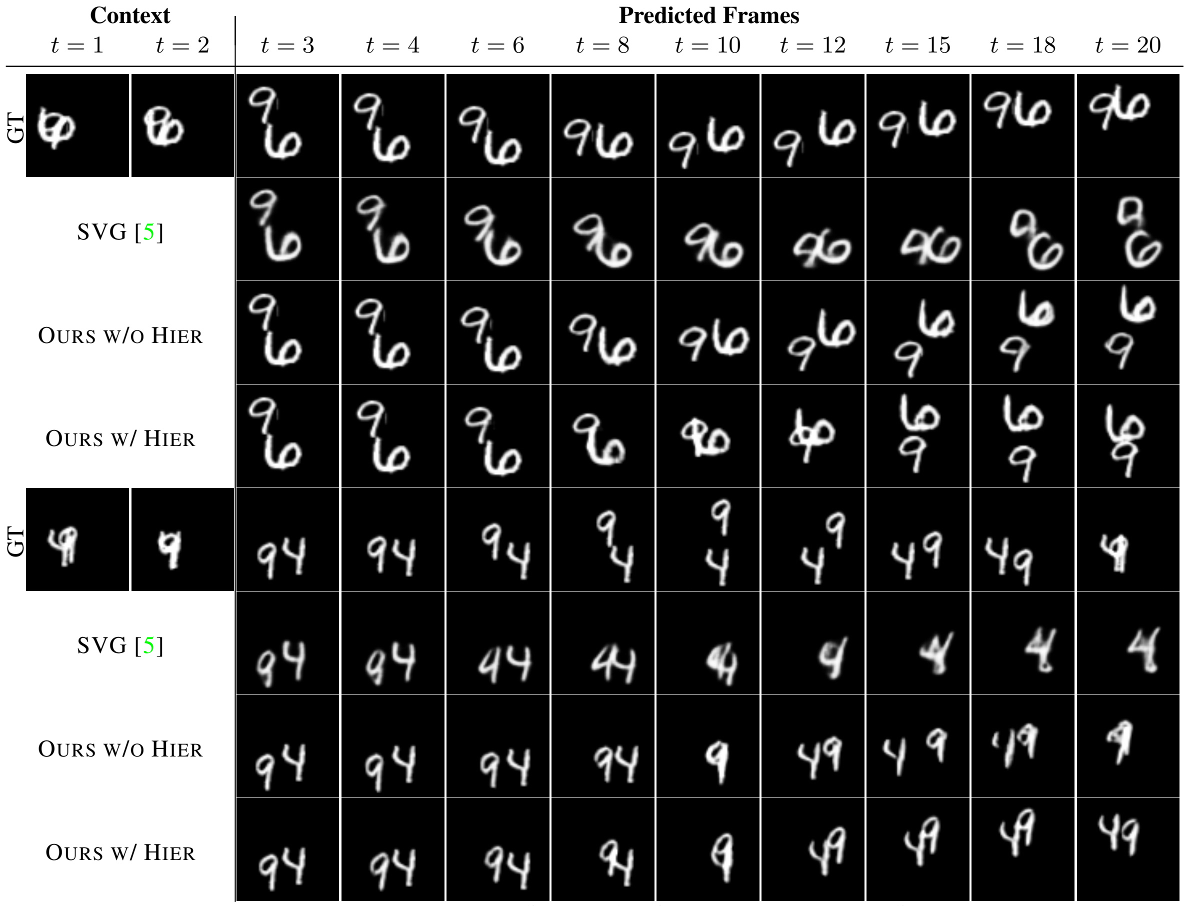


Figure C.3: **Random Test Samples for Stochastic Moving MNIST.**

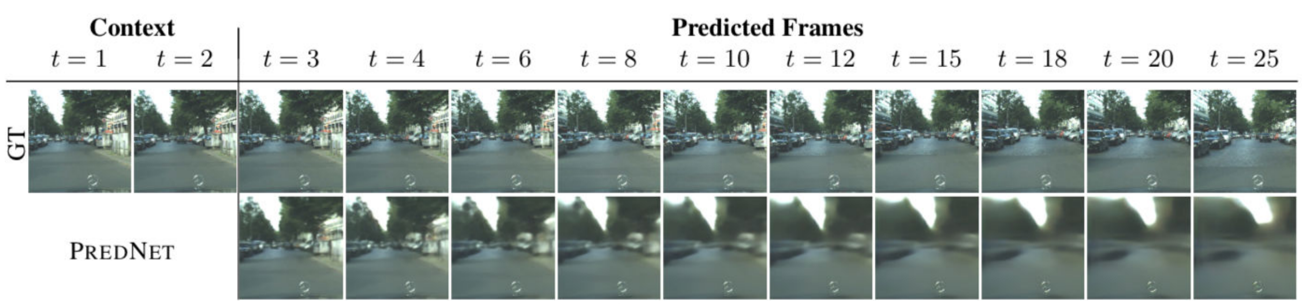


Figure D.1: **A sample from PredNet on Cityscapes.**