A Appendix

We provide some additional information regarding model size, memory requirements, batch size and computation times in Table 8. This is followed by additional results and visualizations for SMMNIST, KTH, BAIR, UCF-101 and CityScapes. Further results, images and videos are also provided in the supplementary material.

A.1 Computational requirements

Table 8: Compute used. "steps" indicates the checkpoint with the best approximate FVD, "GPU hours" is the total training time up to "steps".

Dataset, model	params	CPU mem (GB)	batch size	GPU	GPU mem (GB)	steps	GPU hours
SMMNIST concat	27.9M	3.6	64	Tesla V100	14.5	700000	78.9
SMMNIST spatin	53.9M	3.3	64	RTX 8000	23.4	140000	39.7
KTH concat	62.8M	3.2	64	Tesla V100	21.5	400000	65.7
KTH spatin	367.6M	8.9	64	A100	145.9	340000	45.8
BAIR concat	251.2M	5.1	64	Tesla V100	76.5	450000	78.2
BAIR spatin	328.6M	9.2	64	A100	86.1	390000	50.0
Cityscapes concat	262.1M	6.2	64	Tesla V100	78.2	900000	192.83
Cityscapes spatin	579.1M	8.9	64	A100	101.2	650000	96.0
UCF concat	565.0M	8.9	64	Tesla V100	100.1	900000	183.95
UCF spatin	739.4M	8.9	64	A100	115.2	550000	79.5

A.2 Stochastic Moving MNIST

In Table 9 we provide results for more configurations of our proposed approach on the SMMNIST evaluation. In Figure 5 we provide some visual results for SMMNIST.

Table 9: Results on the SMMNIST evaluation, conditioned on 5 past frames, predicting 10 frames using models trained to predict 5 frames at a time.

SMMNIST [5 $\rightarrow$ 10; trained on 5]	FVD$\downarrow$	PSNR$\uparrow$	SSIM$\uparrow$	LPIPS$\downarrow$	MSE$\downarrow$
MCVD concat	25.63 $\pm$ 0.69	17.22	0.786	0.117	0.024
MCVD concat past-future-mask	20.77 $\pm$ 0.77	16.33	0.753	0.139	0.028
MCVD spatin	23.86 $\pm$ 0.67	17.07	0.785	0.129	0.025
MCVD spatin future-mask	44.14 $\pm$ 1.73	16.31	0.758	0.141	0.027
MCVD spatin past-future-mask	36.12 $\pm$ 0.63	16.15	0.748	0.146	0.027

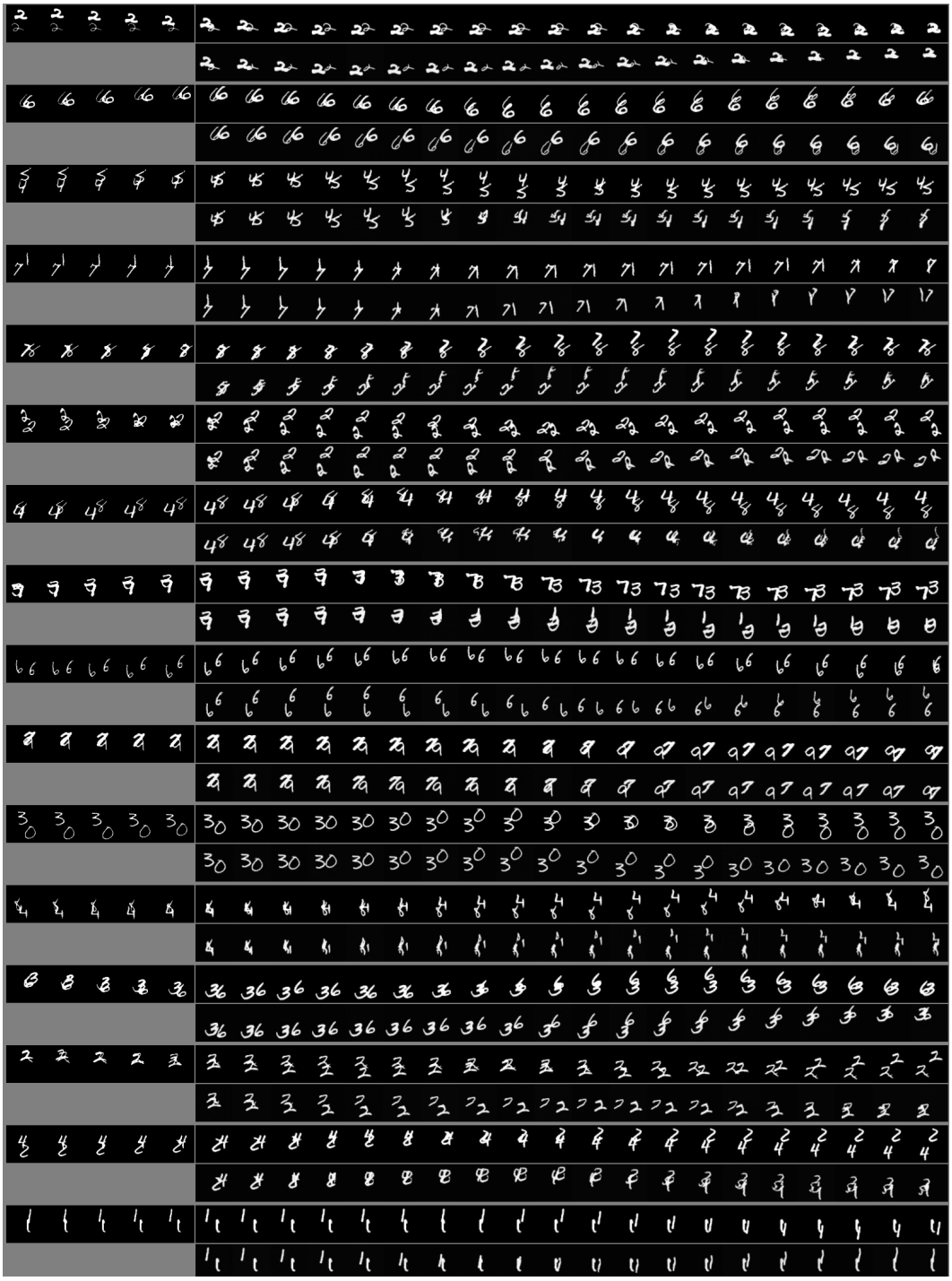

Figure 5: SMMNIST 5 → 10, trained on 5; prediction

A.3 KTH

Table 10: Full table: Results on the KTH evaluation, predicting, 30 and 40 frames using models trained to predict k frames at a time. All models condition on 10 past frames. SSIM numbers are updated from the main text after fixing a bug in the calculation.

KTH [10 $\rightarrow$ $pred$; trained on k]	k	$pred$	FVD↓	PSNR↑	SSIM↑
SV2P [Babaeizadeh et al., 2018b]	10	30	636 ± 1	28.2	0.838
SVG-LP [Denton and Fergus, 2018]	10	30	377 ± 6	28.1	0.844
SAVP [Lee et al., 2018]	10	30	374 ± 3	26.5	0.756
MCVD spatin (Ours)	**5**	30	323 ± 3	27.5	0.835
MCVD concat past-future-mask (Ours)	**5**	30	294.9	24.3	0.746
SLAMP [Akan et al., 2021]	10	30	228 ± 5	29.4	0.865
SRVP [Franceschi et al., 2020]	10	30	222 ± 3	29.7	0.870
Struct-vRNN [Minderer et al., 2019]	10	40	395.0	24.29	0.766
MCVD concat past-future-mask (Ours)	**5**	40	368.4	23.48	0.720
MCVD spatin (Ours)	**5**	40	331.6 ± 5	26.40	0.744
MCVD concat (Ours)	**5**	40	276.6 ± 3	26.20	0.793
SV2P time-invariant [Babaeizadeh et al., 2018b]	10	40	253.5	25.70	0.772
SV2P time-variant [Babaeizadeh et al., 2018b]	10	40	209.5	25.87	0.782
SAVP [Lee et al., 2018]	10	40	183.7	23.79	0.699
SVG-LP [Denton and Fergus, 2018]	10	40	157.9	23.91	0.800
SAVP-VAE [Lee et al., 2018]	10	40	145.7	26.00	0.806
Grid-keypoints [Gao et al., 2021]	10	40	144.2	27.11	0.837

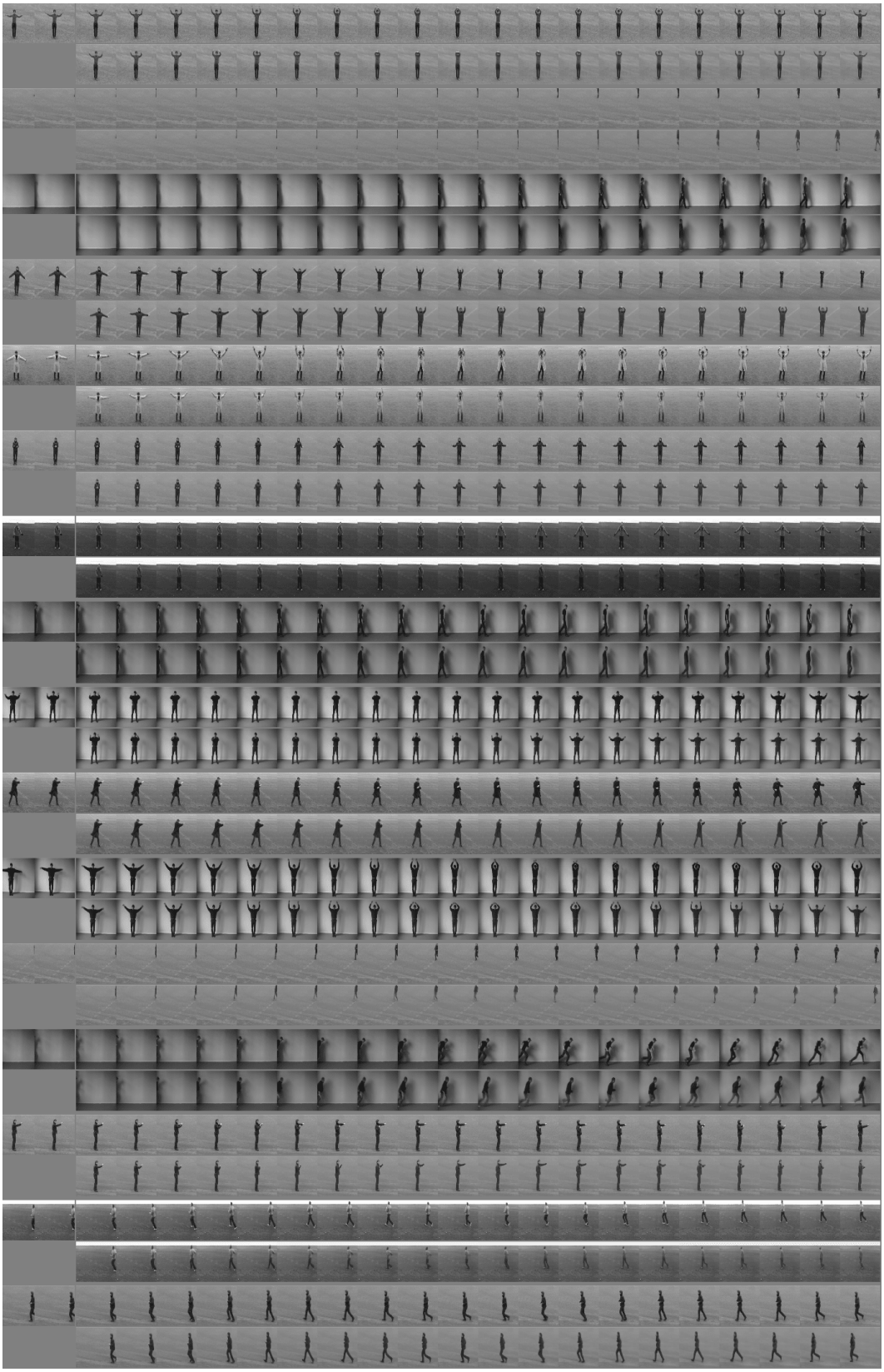

Figure 6: **KTH** 5 $\rightarrow$ 20, trained on 5 (prediction)

A.4 BAIR

Table 11: Full table: Results on the BAIR evaluation conditioning on p past frames and predict pr frames in the future, using models trained to predict k frames at at time.

BAIR [past (p) → pred (pr) ; trained on k]	p	k	pr	FVD↓	PSNR↑	SSIM↑
LVT [Rakhimov et al., 2020]	1	15	15	125.8	–	–
DVD-GAN-FP [Clark et al., 2019]	1	15	15	109.8	–	–
MCVD spatin (Ours)	1	**5**	15	103.8	18.8	0.826
TrIVD-GAN-FP [Luc et al., 2020]	1	15	15	103.3	–	–
VideoGPT [Yan et al., 2021]	1	15	15	103.3	–	–
CCVS [Le Moing et al., 2021]	1	15	15	99.0	–	–
MCVD concat (Ours)	1	**5**	15	98.8	18.8	0.829
MCVD spatin past-mask (Ours)	1	**5**	15	96.5	18.8	0.828
MCVD concat past-mask (Ours)	1	**5**	15	95.6	18.8	0.832
Video Transformer [Weissenborn et al., 2019]	1	15	15	94-96[a]	–	–
FitVid [Babaeizadeh et al., 2021]	1	15	15	93.6	–	–
MCVD concat past-future-mask (Ours)	1	**5**	15	**89.5**	16.9	0.780
SAVP [Lee et al., 2018]	2	14	14	116.4	–	–
MCVD spatin (Ours)	2	**5**	14	94.1	19.1	0.836
MCVD spatin past-mask (Ours)	2	**5**	14	90.5	19.2	0.837
MCVD concat (Ours)	2	**5**	14	90.5	19.1	0.834
MCVD concat past-future-mask (Ours)	2	**5**	14	89.6	17.1	0.787
MCVD concat past-mask (Ours)	2	**5**	14	**87.9**	19.1	0.838
SVG-LP [Akan et al., 2021]	2	10	28	256.6	–	0.816
SVG [Akan et al., 2021].	2	12	28	255.0	18.95	0.8058
SLAMP [Akan et al., 2021]	2	10	28	245.0	19.7	0.818
SRVP [Franceschi et al., 2020]	2	12	28	162.0	19.6	0.820
WAM [Jin et al., 2020]	2	14	28	159.6	21.0	0.844
SAVP [Lee et al., 2018]	2	12	28	152.0	18.44	0.7887
vRNN 1L Castrejón et al. [2019]	2	10	28	149.2	–	0.829
SAVP [Lee et al., 2018]	2	10	28	143.4	–	0.795
Hier-vRNN [Castrejón et al., 2019]	2	10	28	143.4	–	0.822
MCVD spatin (Ours)	2	**5**	28	132.1	17.5	0.779
MCVD spatin past-mask (Ours)	2	**5**	28	127.9	17.7	0.789
MCVD concat (Ours)	2	**5**	28	120.6	17.6	0.785
MCVD concat past-mask (Ours)	2	**5**	28	119.0	17.7	0.797
MCVD concat past-future-mask (Ours)	2	**5**	28	**118.4**	16.2	0.745

[a] 94 on only the first frames, 96 on all subsquences of test frames

Figure 7: **BAIR** 2 $\rightarrow$ 28, trained on 5 (prediction)

A.5 UCF-101

Figure 8: **UCF-101** 4 $\rightarrow$ 16, trained on 4 (**prediction**)

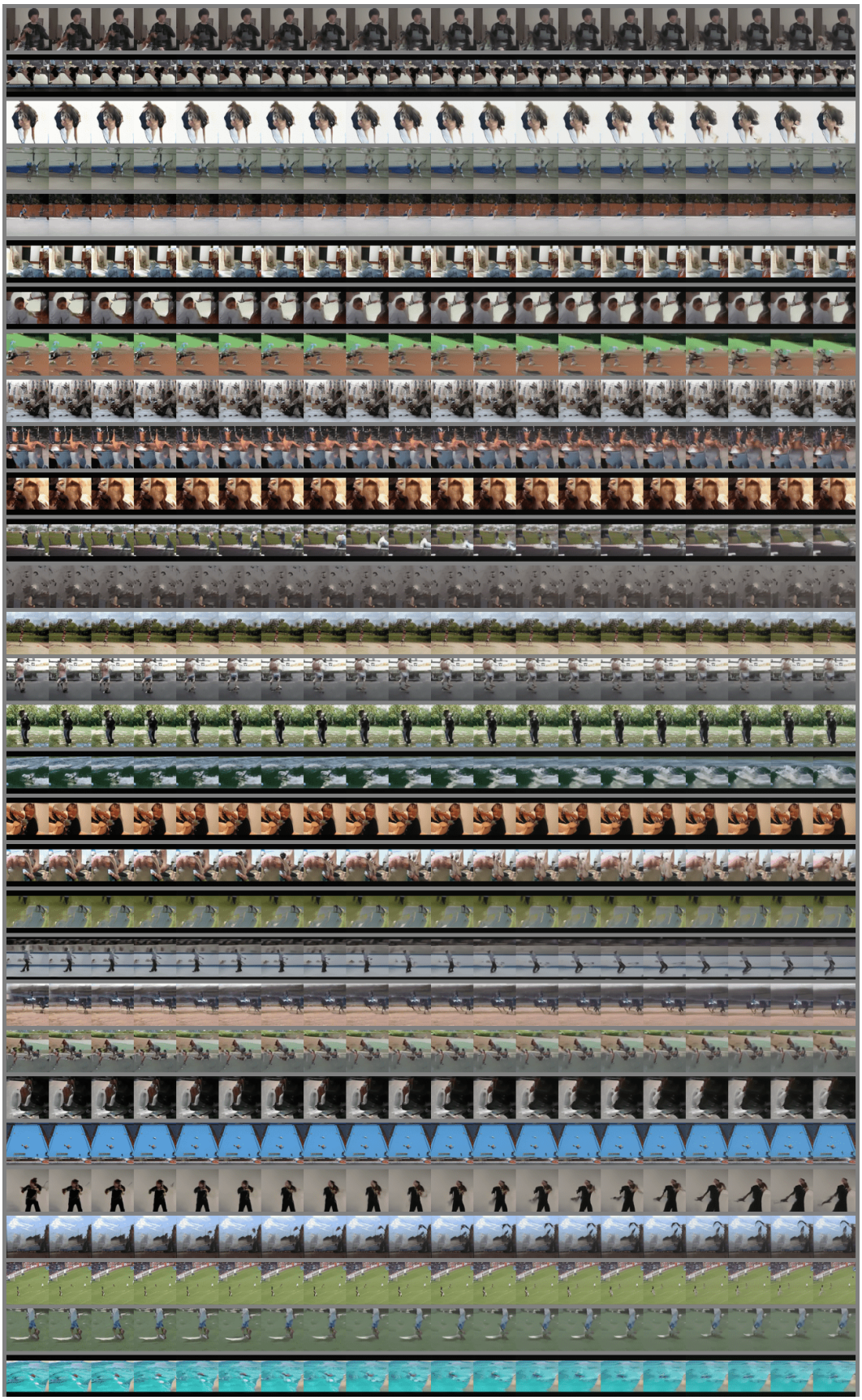

Figure 9: **UCF-101** $0 \rightarrow 4$ (**generation**)

A.6 Cityscapes

Here we provide some examples of future frame prediction for Cityscapes sequences conditioning on two frames and predicting the next 7 frames.

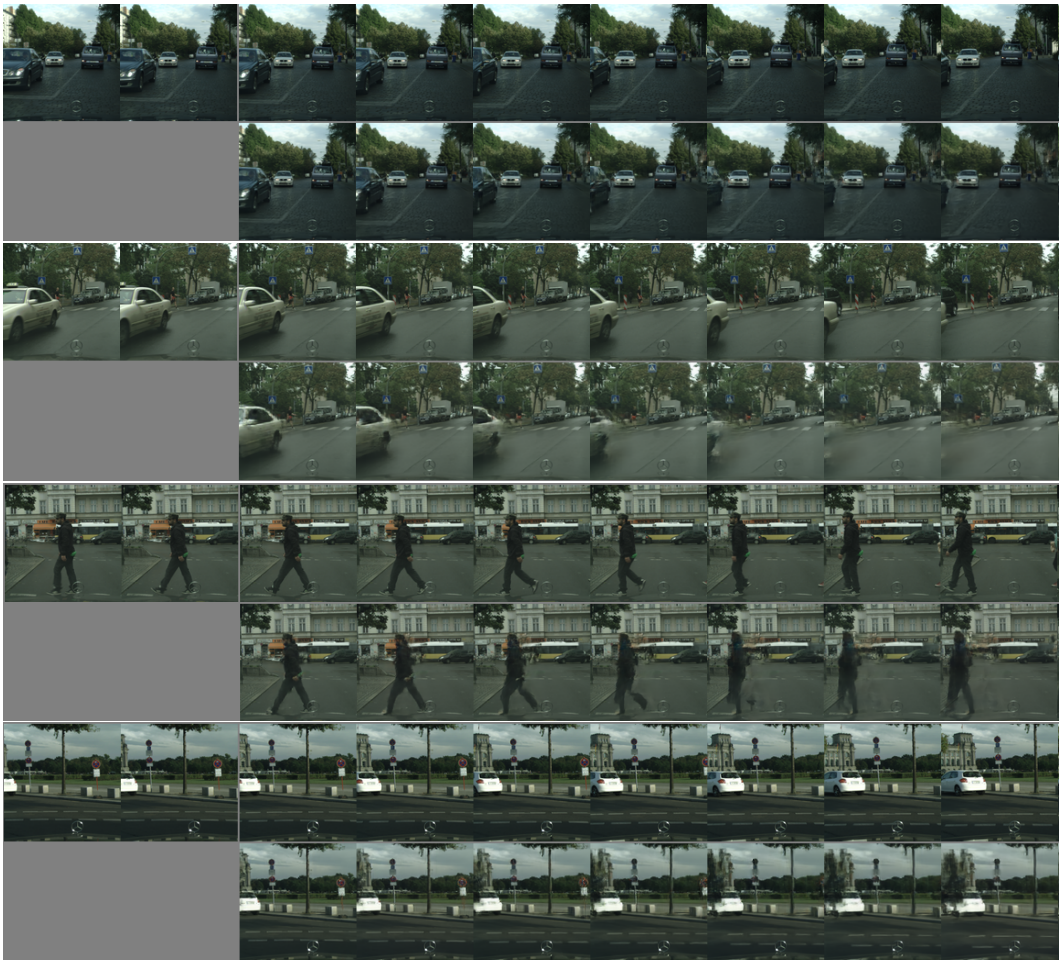

Figure 10: **Cityscapes**: $2 \rightarrow 7$, trained on 5; Conditioning on the two frames in the top left corner of each block of two rows of images, we generate the next 7 frames. The top row is the true frames, bottom row contains the generated frames. We use the **MCVD** concat model variant.