MCVD: Masked Conditional Video Diffusion for Prediction, Generation, and Interpolation

Vikram Voleti*, Alexia Jolicoeur-Martineau*, Christopher Pal

arxiv.org/abs/2205.09853 mask-cond-video-diffusion.github.io

@ NeurIPS 2022

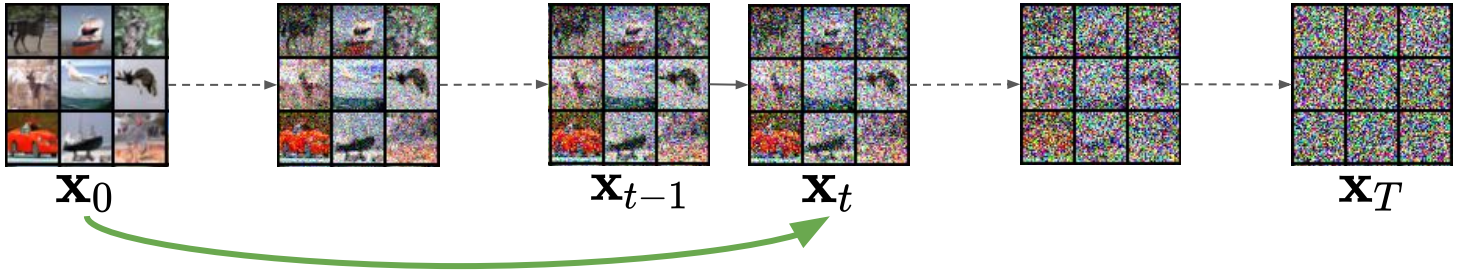
We use score-based denoising diffusion to perform

Video Prediction + Generation + Interpolation

all using the same model!

MCVD: Masked Conditional Video Diffusion

Denoising Diffusion Probabilistic Models (DDPM)



$$q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \quad ; 1 > \bar{\alpha}_1 > \cdots > \bar{\alpha}_t > \cdots > \bar{\alpha}_T > 0$$

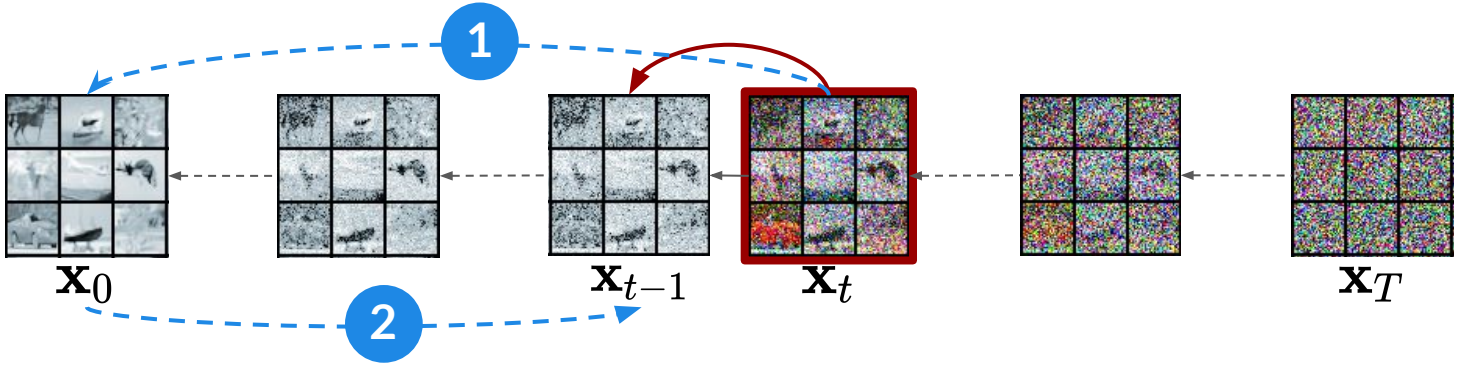
$$\implies \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \quad ; \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Noise matching: predict ϵ from $\mathbf{x}_t$

$$\mathbb{E}_{t,\ \mathbf{x}_0 \sim p_{\text{data}},\ \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})} \left|\left|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\right|\right|_2^2$$

arxiv.org/abs/2006.11239

DDPM : Generation



$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

for $t = T \rightarrow 1$:

1 $\hat{\mathbf{x}}_0 = (\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(\mathbf{x}_t, t))/\sqrt{\bar{\alpha}_t}$

2 $\mathbf{x}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}\,\mathbf{z}_t$

$(\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}))$

arxiv.org/abs/2006.11239

Solving video tasks:

Video Prediction + Generation + Interpolation

using denoising diffusion

arxiv.org/abs/2205.09853

MCVD: Masked Conditional Video Diffusion

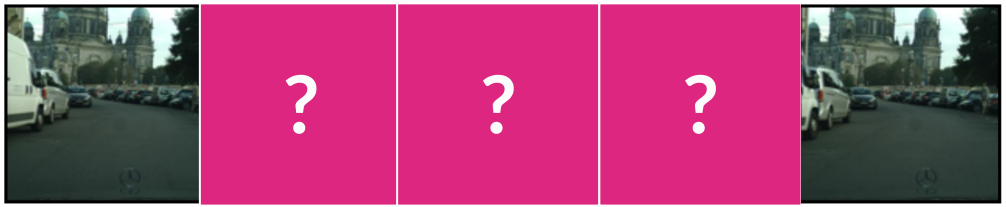
Mila

Real data

Video Prediction

Video Generation

Video Interpolation



arxiv.org/abs/2205.09853

MCVD: Masked Conditional Video Diffusion

Mila

p past frames: $\mathbf{p} = \{\mathbf{p}^i\}_{i=1}^{p}$ | k current frames: $\mathbf{x}_0 = \{\mathbf{x_0}^i\}_{i=1}^{k}$ | f future frames: $\mathbf{f} = \{\mathbf{f}^i\}_{i=1}^{f}$ | $(\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon)$

- **Video Prediction:**

$$\mathbb{E}_{t,\ [\mathbf{p},\mathbf{x}_0]\sim p_{\text{data}},\ \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left|\left|\epsilon - \epsilon_\theta(\mathbf{x}_t \mid \mathbf{p}, t)\right|\right|_2^2$$

Can we do better?

- **Video Generation:**

$$\mathbb{E}_{t,\ \mathbf{x}_0\sim p_{\text{data}},\ \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left|\left|\epsilon - \epsilon_\theta(\mathbf{x}_t \mid t)\right|\right|_2^2$$

- **Video Interpolation:**

$$\mathbb{E}_{t,\ [\mathbf{p},\mathbf{x}_0,\mathbf{f}]\sim p_{\text{data}},\ \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left|\left|\epsilon - \epsilon_\theta(\mathbf{x}_t \mid \mathbf{p}, \mathbf{f}, t)\right|\right|_2^2$$

mask-cond-video-diffusion.github.io, arxiv.org/abs/2205.09853

MCVD: Masked Conditional Video Diffusion

Mila

p past frames: $\mathbf{p} = \{\mathbf{p}^i\}_{i=1}^{p}$ | k current frames: $\mathbf{x}_0 = \{\mathbf{x_0}^i\}_{i=1}^{k}$ | f future frames: $\mathbf{f} = \{\mathbf{f}^i\}_{i=1}^{f}$ $\quad (\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon)$

- **Video Prediction:**

$$\mathbb{E}_{t,\ [\mathbf{p},\mathbf{x}_0]\sim p_{\text{data}},\ \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left|\left|\epsilon - \epsilon_\theta(\mathbf{x}_t \mid \mathbf{p}, t)\right|\right|_2^2$$

Random masking!

- **Video Prediction + Generation:**

$$\mathbb{E}_{t,\ [\mathbf{p},\mathbf{x}_0]\sim p_{\text{data}},\ \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\ m_p\sim\mathcal{B}(p_{\text{mask}})} \left|\left|\epsilon - \epsilon_\theta(\mathbf{x}_t \mid m_p\mathbf{p}, t)\right|\right|_2^2$$

- **Video Prediction + Generation + Interpolation:**

$$\mathbb{E}_{t,\ [\mathbf{p},\mathbf{x}_0,\mathbf{f}]\sim p_{\text{data}},\ \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\ (m_p,m_f)\sim\mathcal{B}(p_{\text{mask}})} \left|\left|\epsilon - \epsilon_\theta(\mathbf{x}_t \mid m_p\mathbf{p}, m_f\mathbf{f}, t)\right|\right|_2^2$$

mask-cond-video-diffusion.github.io, arxiv.org/abs/2205.09853

MCVD: Masked Conditional Video Diffusion

Mila

Video Prediction

Video Prediction + Generation + Interpolation

mask-cond-video-diffusion.github.io, arxiv.org/abs/2205.09853

MCVD: Masked Conditional Video Diffusion

Mila

Video Generation

Video Prediction + Generation + Interpolation

mask-cond-video-diffusion.github.io, arxiv.org/abs/2205.09853

MCVD: Masked Conditional Video Diffusion

Mila

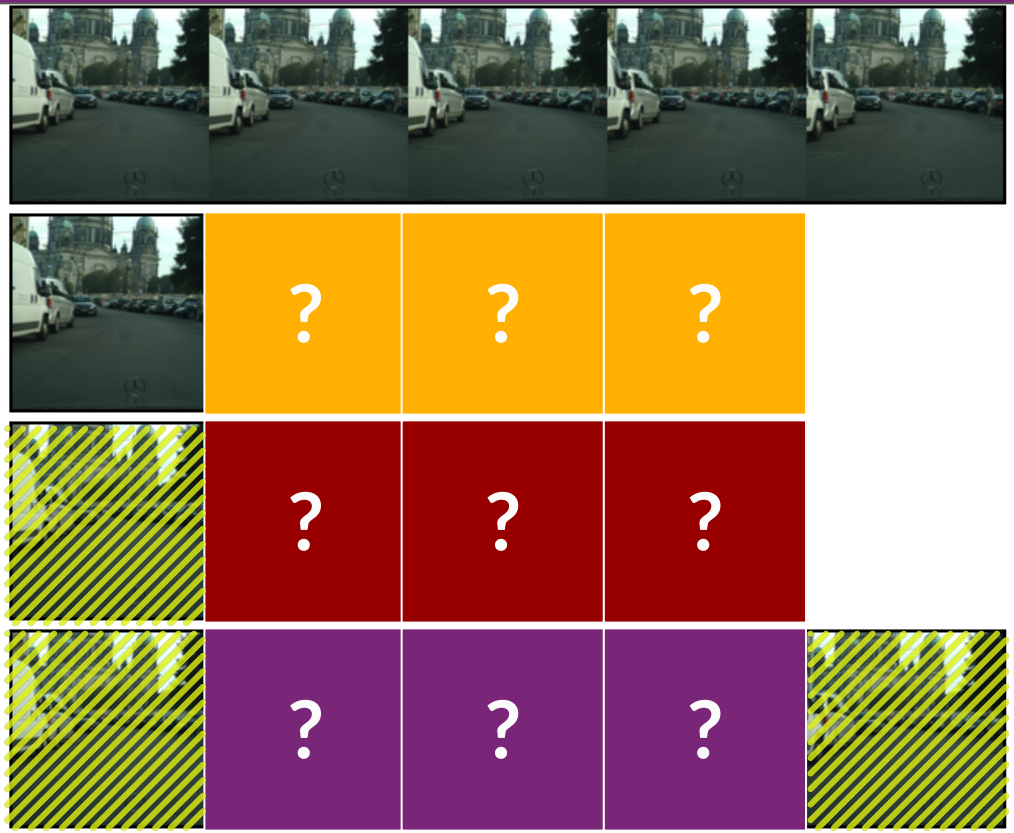
Video Interpolation

Video Prediction + Generation + Interpolation

mask-cond-video-diffusion.github.io, arxiv.org/abs/2205.09853

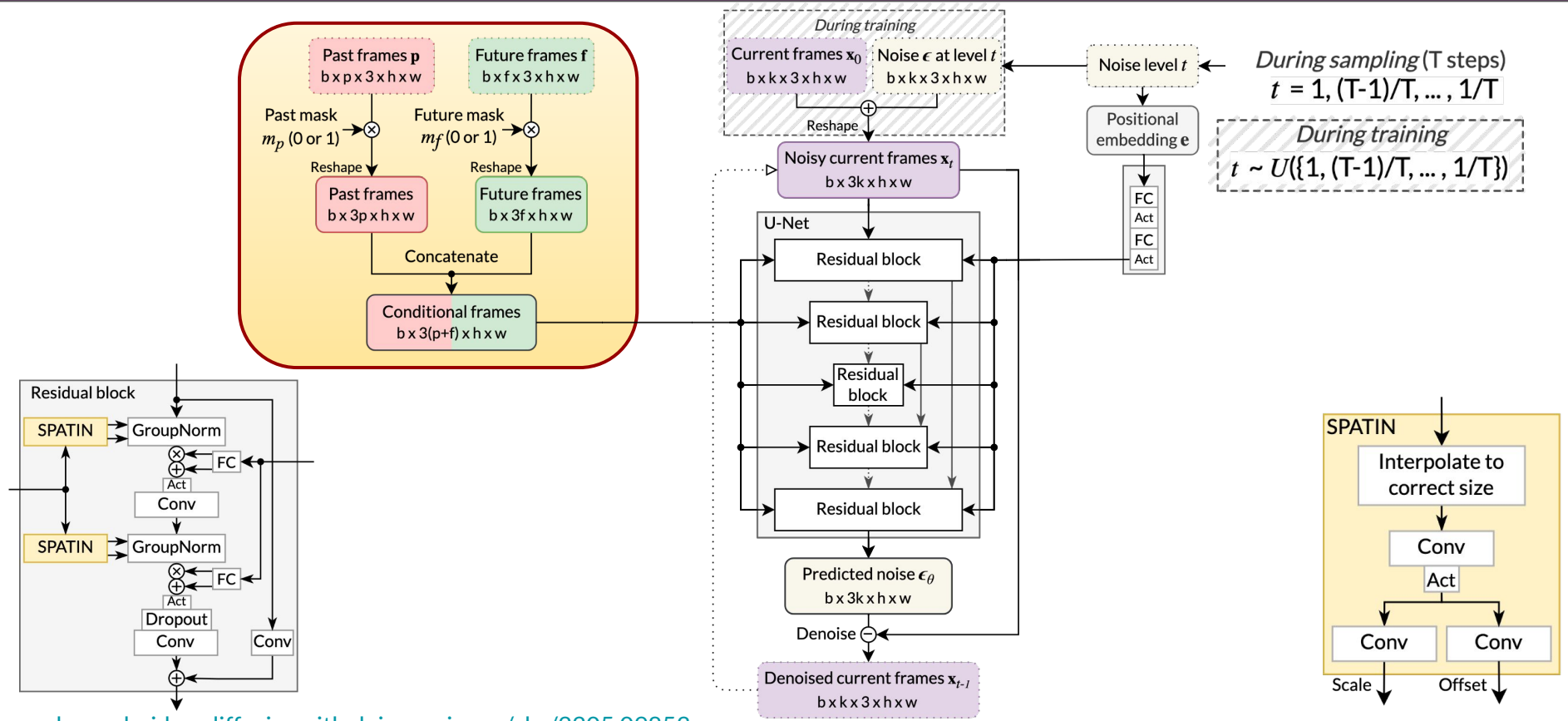
MCVD: Masked Conditional Video Diffusion

Mila



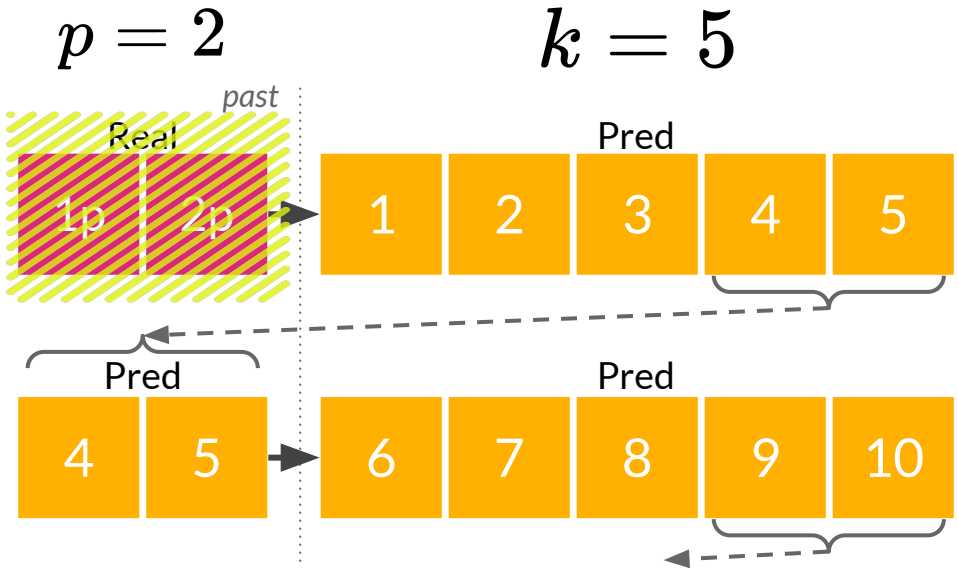
mask-cond-video-diffusion.github.io, arxiv.org/abs/2205.09853

MCVD: Masked Conditional Video Diffusion



mask-cond-video-diffusion.github.io, arxiv.org/abs/2205.09853

Block-autoregressive generation:



mask-cond-video-diffusion.github.io, arxiv.org/abs/2205.09853

MCVD: Masked Conditional Video Diffusion

Mila

(128x128)

Cityscapes [2 → 28; trained on k]	k	FVD↓	LPIPS↓
SVG-LP Denton and Fergus [2018]	10	1300.26	0.549 ± 0.06
vRNN 1L Castrejón et al. [2019]	10	682.08	0.304 ± 0.10
Hier-vRNN Castrejón et al. [2019]	10	567.51	0.264 ± 0.07
GHVAE Wu et al. [2021]	10	418.00	0.193 ± 0.014
MCVD spatin (Ours)	**5**	184.81	0.121 ± 0.05
MCVD concat (Ours)	**5**	**141.31**	**0.112** ± 0.05

(64x64)

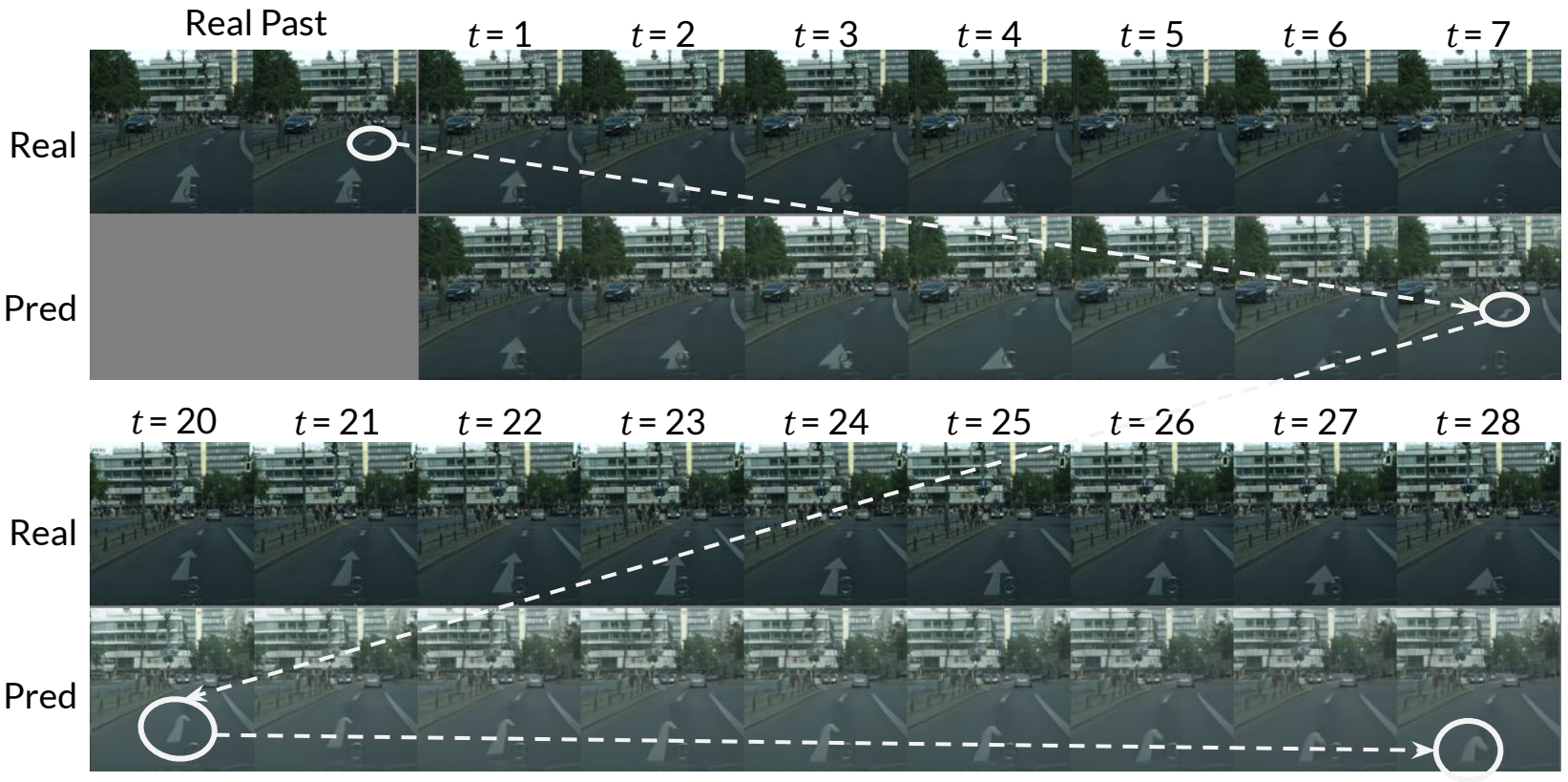
SMMNIST [5 → 10; trained on k]	k	FVD↓	SSIM↑
SVG [Denton and Fergus, 2018]	10	90.81	0.688
vRNN 1L [Castrejón et al., 2019]	10	63.81	0.763
Hier-vRNN [Castrejón et al., 2019]	10	57.17	0.760
MCVD concat (Ours)	**5**	25.63	**0.786**
MCVD spatin (Ours)	**5**	23.86	0.780
MCVD concat past-future-mask	**5**	**20.77**	0.753

mask-cond-video-diffusion.github.io, arxiv.org/abs/2205.09853

(64x64)

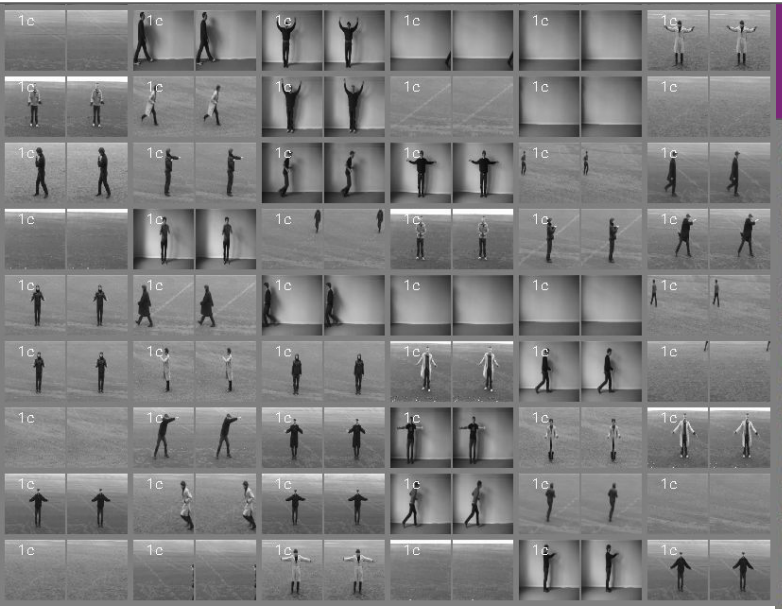
BAIR [past p → $pred$; trained on k]	p	k	$pred$	FVD↓
LVT [Rakhimov et al., 2020]	1	15	15	125.8
DVD-GAN-FP [Clark et al., 2019]	1	15	15	109.8
MCVD spatin (Ours)	1	**5**	15	103.8
TrIVD-GAN-FP [Luc et al., 2020]	1	15	15	103.3
VideoGPT [Yan et al., 2021]	1	15	15	103.3
CCVS [Le Moing et al., 2021]	1	15	15	99.0
MCVD concat (Ours)	1	**5**	15	98.8
MCVD spatin past-mask (Ours)	1	**5**	15	96.5
MCVD concat past-mask (Ours)	1	**5**	15	**95.6**
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**
SAVP [Lee et al., 2018]	2	14	14	116.4
MCVD spatin (Ours)	2	**5**	14	94.1
MCVD spatin past-mask (Ours)	2	**5**	14	90.5
MCVD concat (Ours)	2	**5**	14	90.5
MCVD concat past-future-mask (Ours)	2	**5**	14	89.6
MCVD concat past-mask (Ours)	2	**5**	14	**87.9**
SVG-LP [Akan et al., 2021]	2	10	28	256.6
SLAMP [Akan et al., 2021]	2	10	28	245.0
SAVP [Lee et al., 2018]	2	10	28	143.4
Hier-vRNN [Castrejón et al., 2019]	2	10	28	143.4
MCVD spatin (Ours)	2	**5**	28	132.1
MCVD spatin past-mask (Ours)	2	**5**	28	127.9
MCVD concat (Ours)	2	**5**	28	120.6
MCVD concat past-mask (Ours)	2	**5**	28	119.0
MCVD concat past-future-mask (Ours)	2	**5**	28	**118.4**

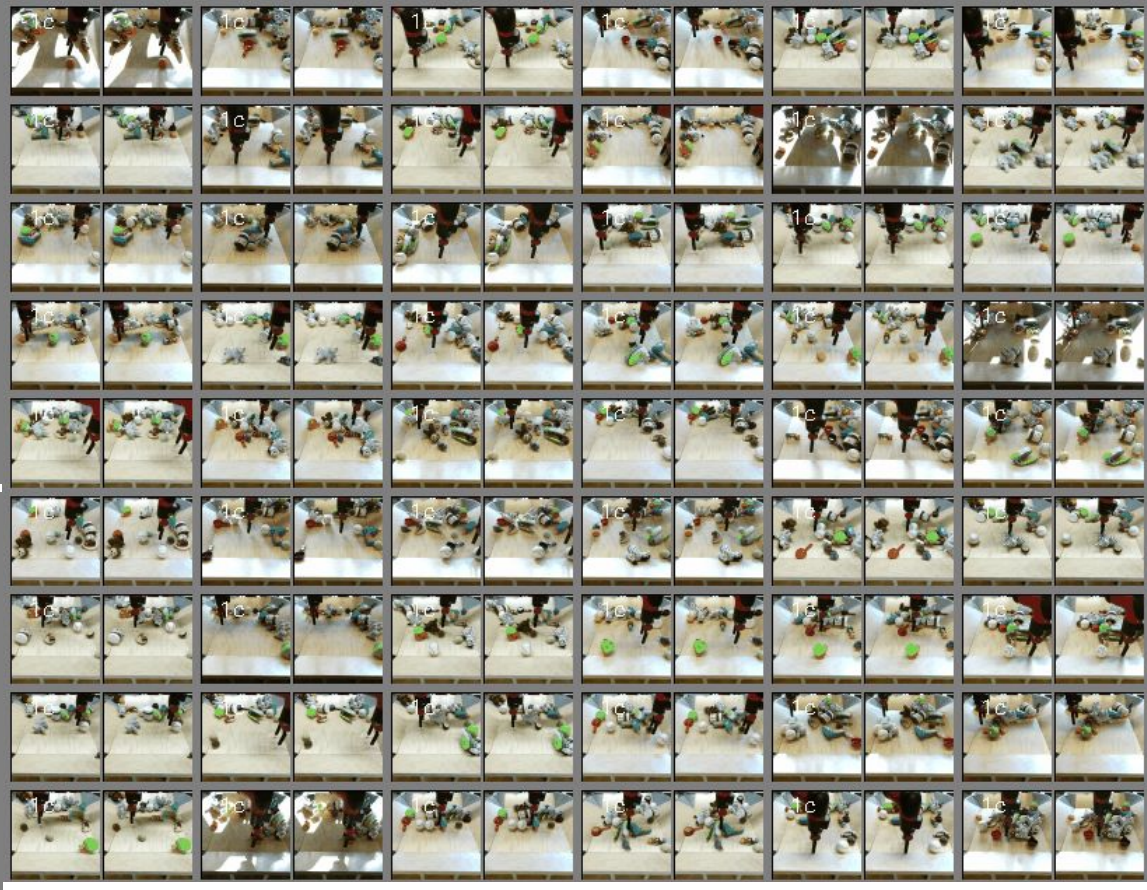
MCVD: Masked Conditional Video Diffusion



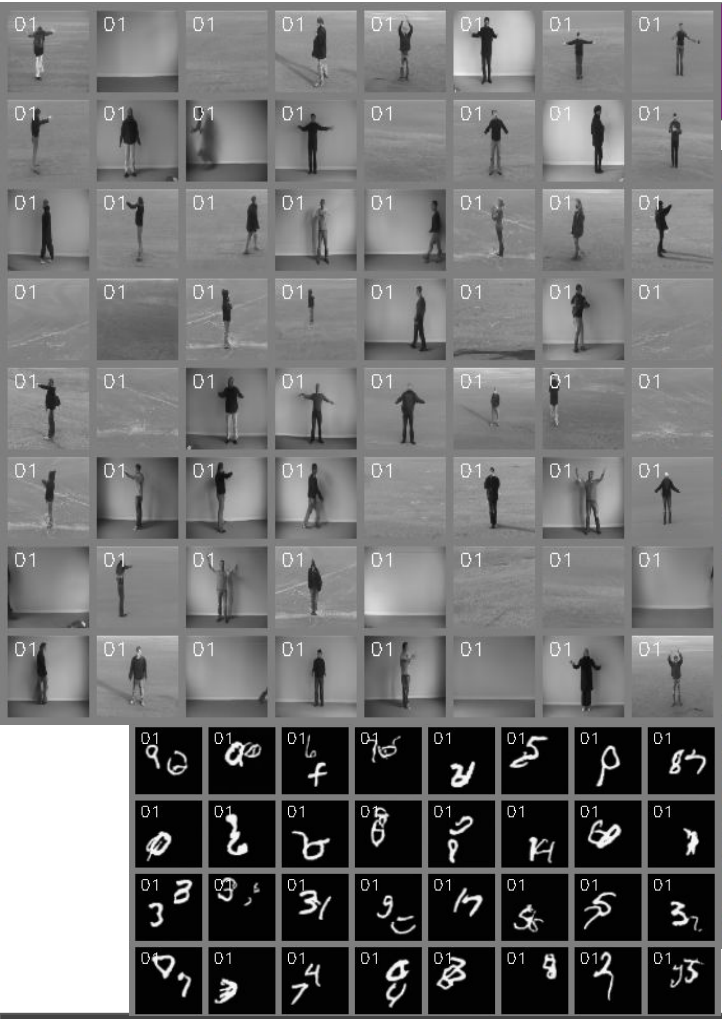
mask-cond-video-diffusion.github.io

Interpolation





Generation



Prediction

We use score-based denoising diffusion to perform

Video Prediction + Generation + Interpolation

all using the same model!

arxiv.org/abs/2205.09853

MCVD: Masked Conditional Video Diffusion for Video Prediction, Generation, and Interpolation

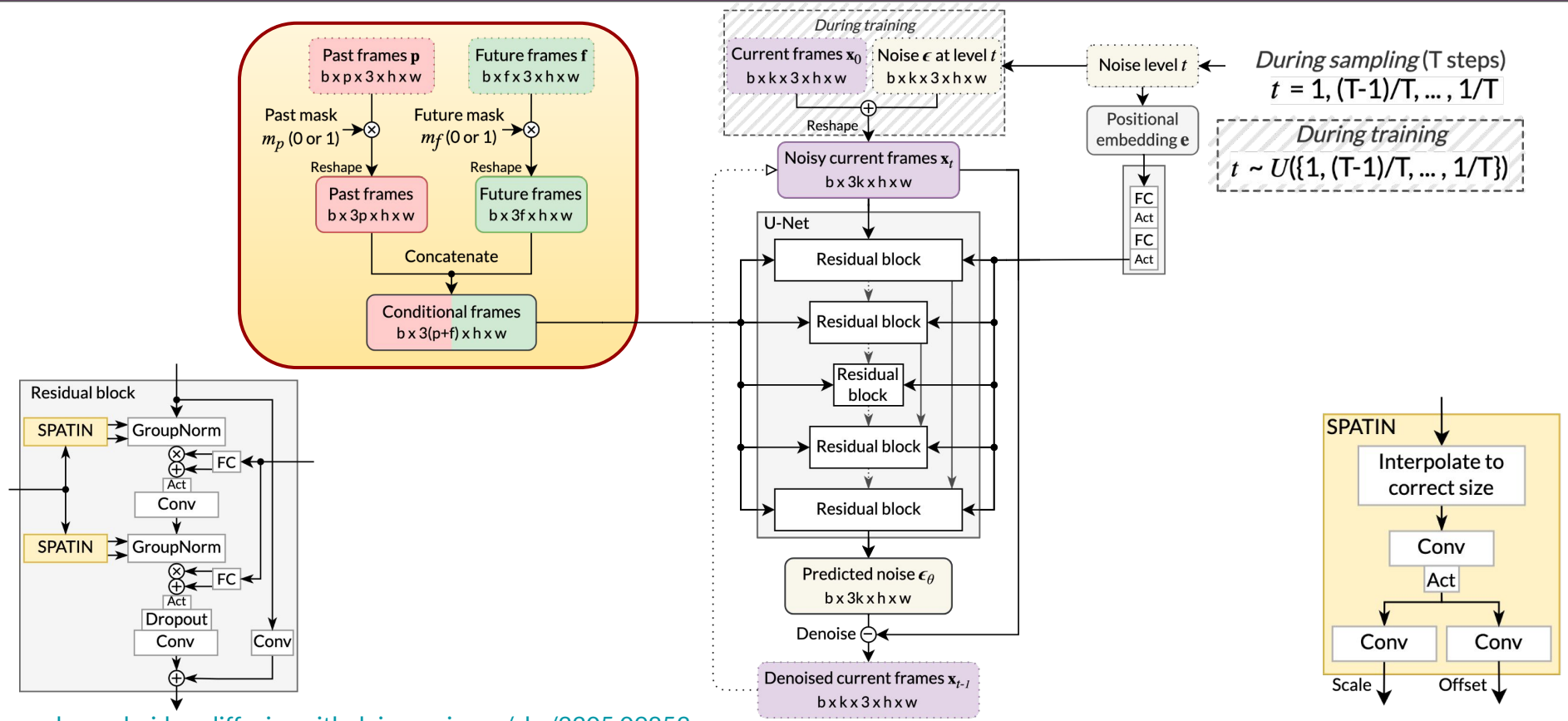
Mila

mask-cond-video-diffusion.github.io

NeurIPS 2022

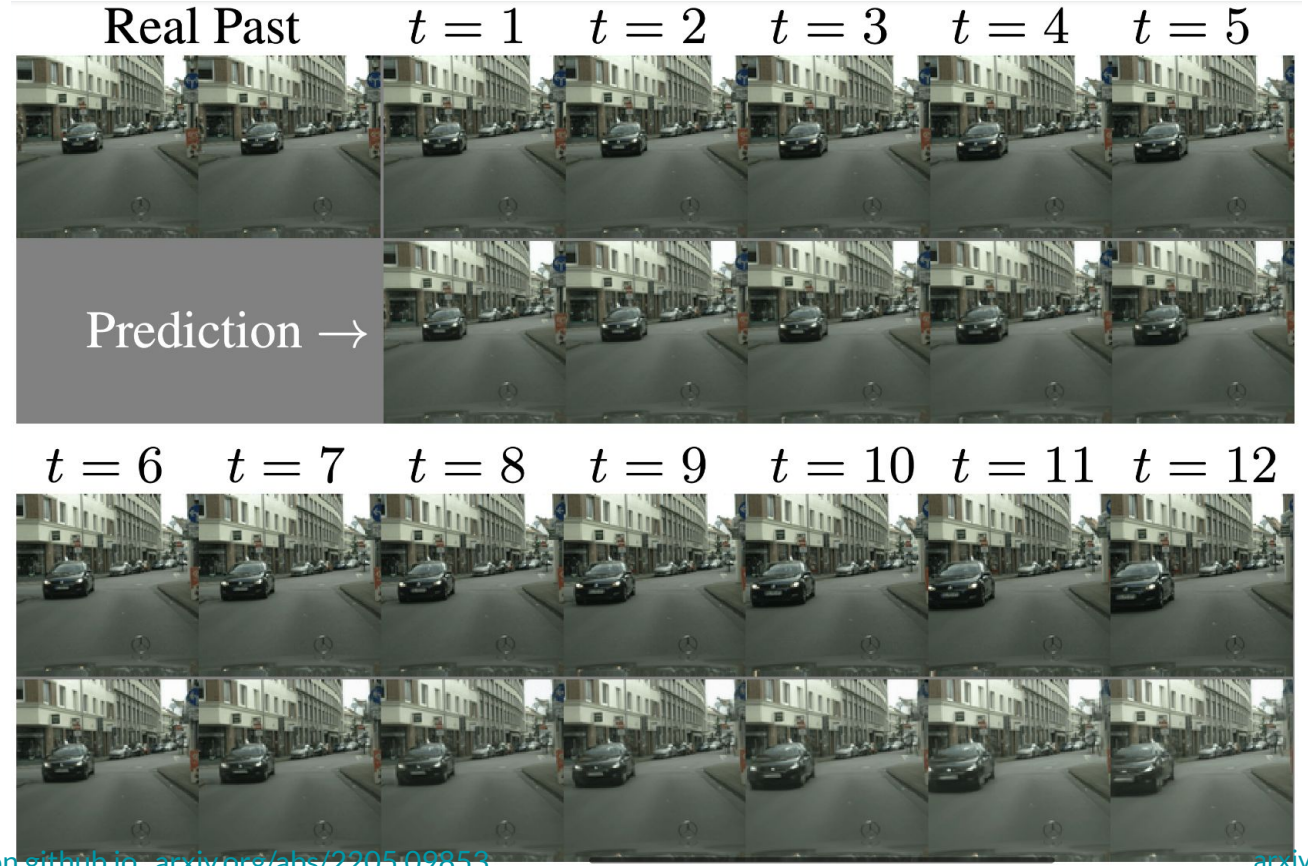
Vikram Voleti*, Alexia Jolicoeur-Martineau*, Christopher Pal

MCVD: Masked Conditional Video Diffusion



mask-cond-video-diffusion.github.io, arxiv.org/abs/2205.09853

MCVD: Masked Conditional Video Diffusion



mask-cond-video-diffusion.github.io, arxiv.org/abs/2205.09853

arxiv.org/abs/2205.09853