CVPR SEATTLE, WA JUNE 17-21, 2024

ExtDM: Distribution Extrapolation Diffusion Model for Video Prediction

Zhicheng Zhang[1,2,†] Junyao Hu[1,2,†] Wentao Cheng[1,‡] Danda Paudel[3,4] Jufeng Yang[1,2]

[1] VCIP & TMCC & DISSec, College of Computer Science, Nankai University
[2] Nankai International Advanced Research Institute (SHENZHEN· FUTIAN)
[3] Computer Vision Lab, ETH Zurich [4] INSAIT, Sofia University

https://cv.nankai.edu.cn
https://github.com/nku-zhichengzhang/ExtDM

Project Code

Outline

- **Introduction**
- **Rethinking Previous Works**
- **ExtDM Architecture**
- **Experimental Results**
- **Conclusion**

Introduction

Autonomous Driving

Sport Events

- **Video Prediction**
  - **Definition**

    It aims to capture the dynamic change from present x_c to future x_p.
  - **Difference with Video Generation**

    building on existing video sequences *v.s.* creating from scratch
  - **Application**

    Autonomous driving, sport events, video understanding, *etc.*

Introduction

Prediction Performance of MCVD

Methods	*cond*=10, *pred* = 40				FPS↑
	SSIM↑	PSNR↑	LPIPS↓	FVD↓	
MCVD-c	0.793	26.20	0.124	276.6	6.35
MCVD-cpf	0.720	23.48	0.173	368.4	6.38
MCVD-s	0.744	26.40	0.115	331.6	2.29

Inference quality and speed of MCVD

- **Video Prediction**
  - **Challenges**
    - **Uncertainty and Complexity**

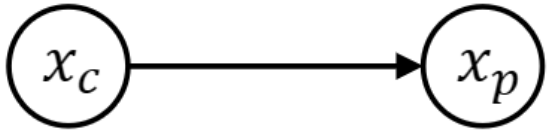
      especially in long-term video prediction
    - **Modeling of Temporal Change**

      including dynamic variation and static background processing
    - **Effectiveness and Usability**

      Trade-off between training computing cost and inference speed

Voleti V, Jolicoeur-Martineau A, Pal C. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation[C]. NeurIPS, 2022.

Rethinking Previous Works



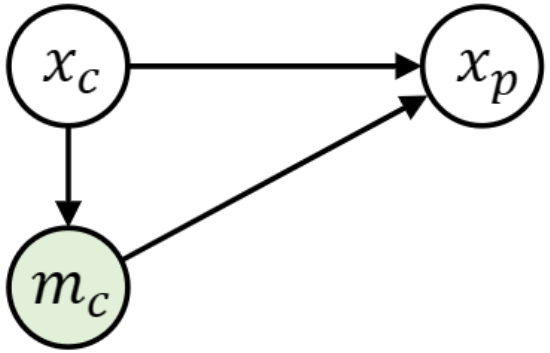
$$p(x_p|x_c)$$

Direct Method

- only **RGB**
- difficult to solve **complexity** in probability estimation

  SRVP(ICML20)

  SimVP(CVPR22)



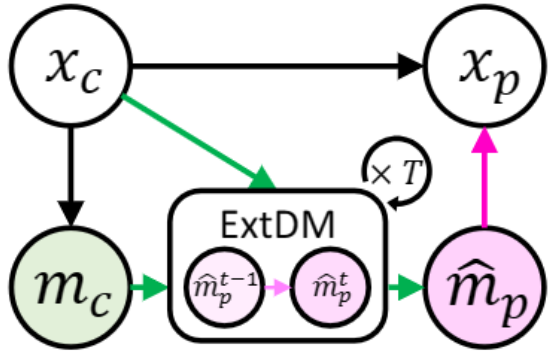
$$p(x_p|x_c, m_c)$$

In-context Learning Method

- **RGB** + **motion** (**implicit** cues)
- lack accuracy for **longer** time
- **counterfactual** results like fading, deformation, etc.

  MCNet(ICLR17)

  MOSO(CVPR23)

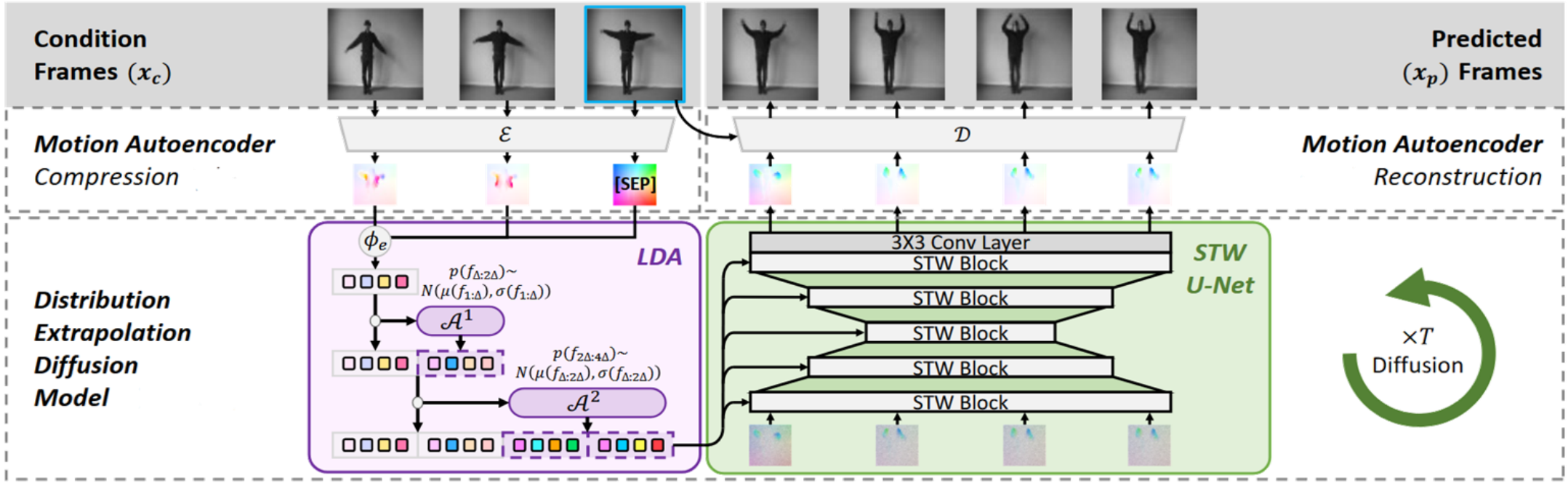


$$p\left(x_p|x_c, \zeta(x_c, m_c)\right)$$

Extrapolation Method (Ours)

- **RGB** + **motion** (**explicit** cues)
- **Extrapolate** present deterministic motion cues into the future ones

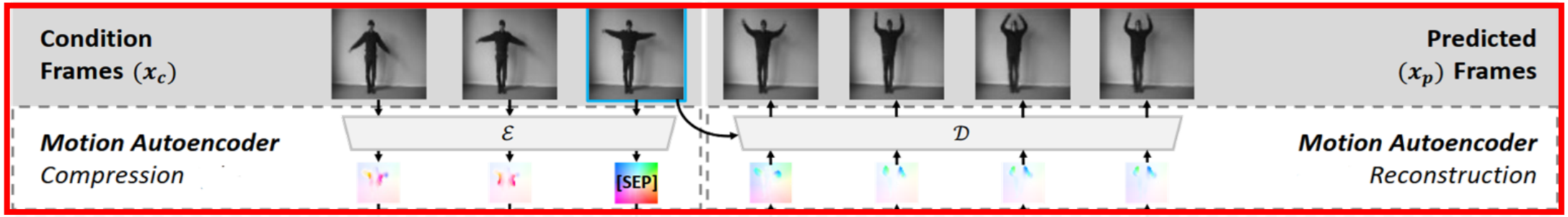
ExtDM Architecture



✓ Contributions

- A **distribution extrapolation** DM that predict future frames.
- An efficient VP method includes **compression and reconstruction**, which can create multiple tailored proposals for stochastic events by imitating motion cues.
- Effectiveness for **short/long-term** videos in 5 video prediction datasets.

ExtDM Architecture

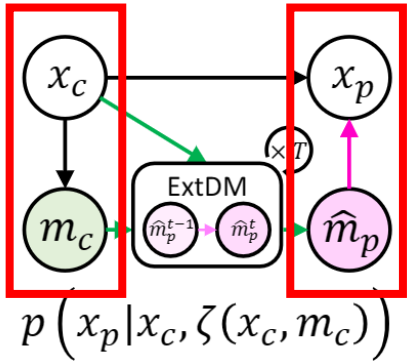


Compression

$$\boldsymbol{m}_c = \left\{ \boldsymbol{m}_i \in \mathbb{R}^{3hw} \mid \boldsymbol{m}_i = \mathcal{E}(\boldsymbol{x}_i, \boldsymbol{x}_u) = \begin{bmatrix} w_i \\ o_i \end{bmatrix} \right\}.$$

Reconstruction

$$\boldsymbol{x}_p = \{ \boldsymbol{x}_j \in \mathbb{R}^{3HW} \mid \boldsymbol{x}_j = \mathcal{D}(\hat{\boldsymbol{m}}_j, \boldsymbol{x}_u) = \mathcal{G}(o_j \odot \mathcal{W}(\boldsymbol{z}_u, w_j)) \}$$



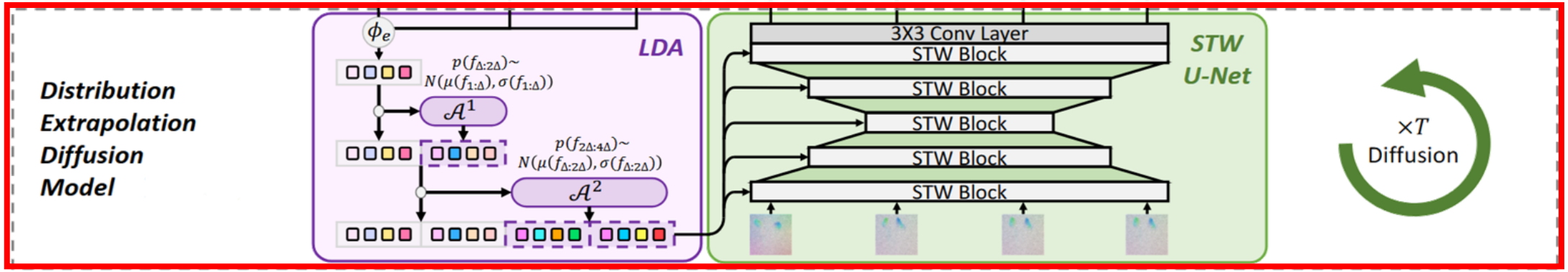
Two Mapping Functions

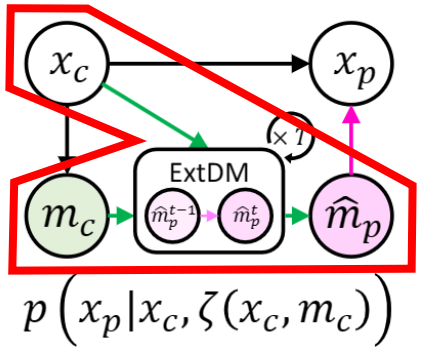
Step 1: $\boldsymbol{x}_c \rightarrow \boldsymbol{m}_c$ & $\hat{\boldsymbol{m}}_p \rightarrow \boldsymbol{x}_p$

- Motion Autoencoder Compression & Reconstruction

ExtDM Architecture

$$\mathcal{L}_{ext} = \mathbb{E}_{t\sim\mathcal{U}(0,1),\hat{\boldsymbol{m}}_p^0\sim q(\hat{\boldsymbol{m}}_p^0),\epsilon\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[||\epsilon - \epsilon_\theta(\hat{\boldsymbol{m}}_p^t, t, c)||^2\right]$$



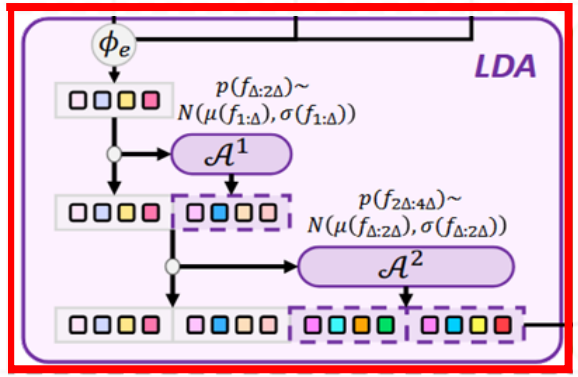


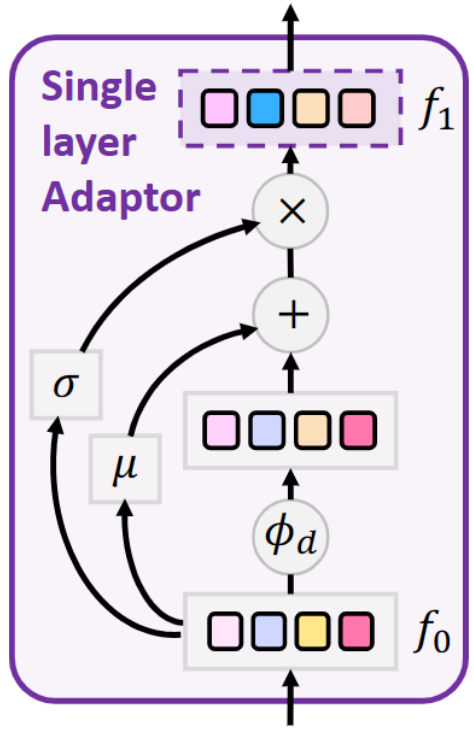
Two Mapping Functions

Step 2: $\boldsymbol{m}_c, \boldsymbol{x}_c \rightarrow \hat{\boldsymbol{m}}_p$

- Distribution Extrapolation Diffusion Model

ExtDM Architecture





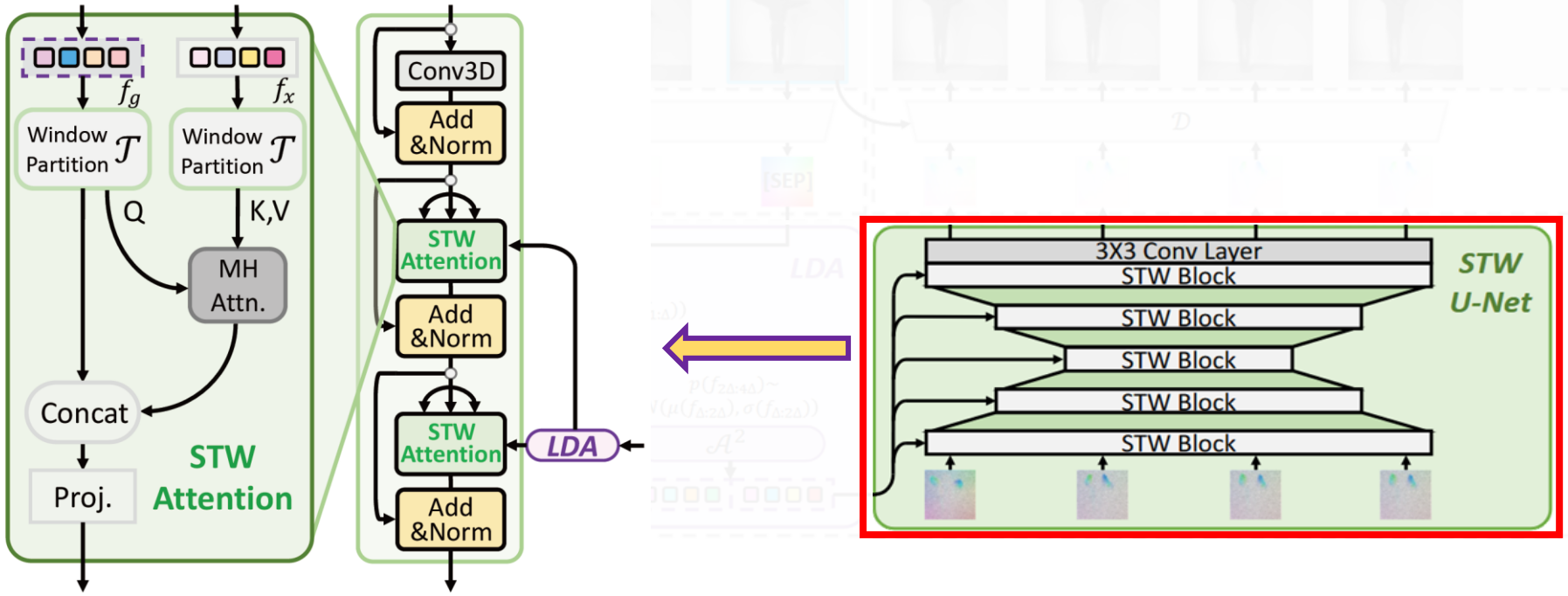
Layered Distribution Adaptor

- estimate distribution params
- inference using distribution sampling

$$f_{1:\Delta} = \phi_e(f_c),$$
$$\widehat{f}_{1:2^l\Delta} = (f_{1:2^{l-1}\Delta}, \mathcal{A}^{(l)}(f_{1:2^{l-1}\Delta})),$$
$$f_p = (\widehat{f}_{1:\Delta}, \ldots, \widehat{f}_{2^{L-1}\Delta:2^L\Delta}).$$

$$f_b = \mathcal{A}(f_a)$$
$$= (\sigma(f_a) + \sigma')\phi_d\left(\frac{f_a - \mu(f_a)}{\sigma(f_a)}\right) + \mu(f_a) + \mu'$$

ExtDM Architecture

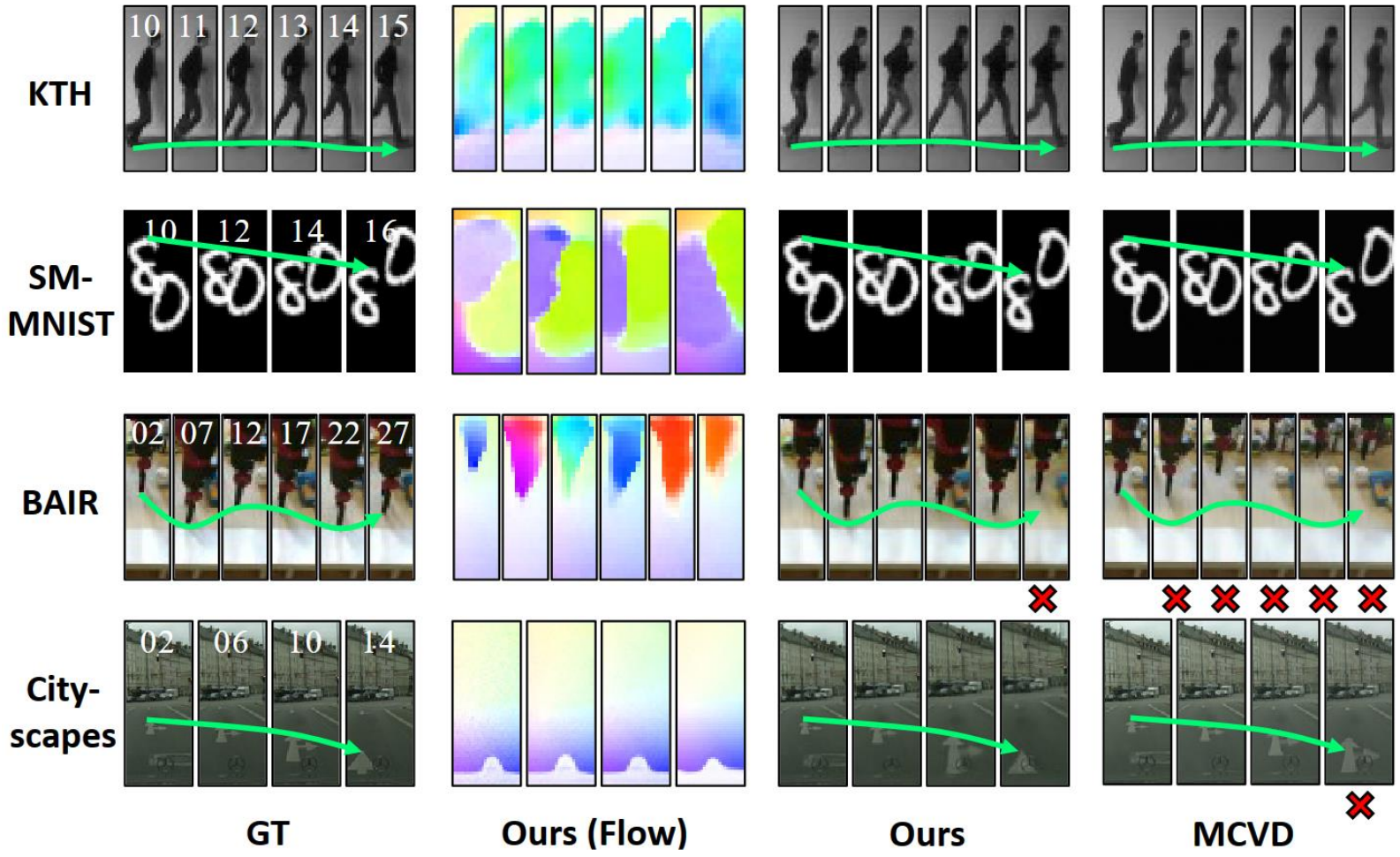


Spatiotemporal Window U-Net

- exploit the spatiotemporal coherence interaction via jointly conducting strided and grid window

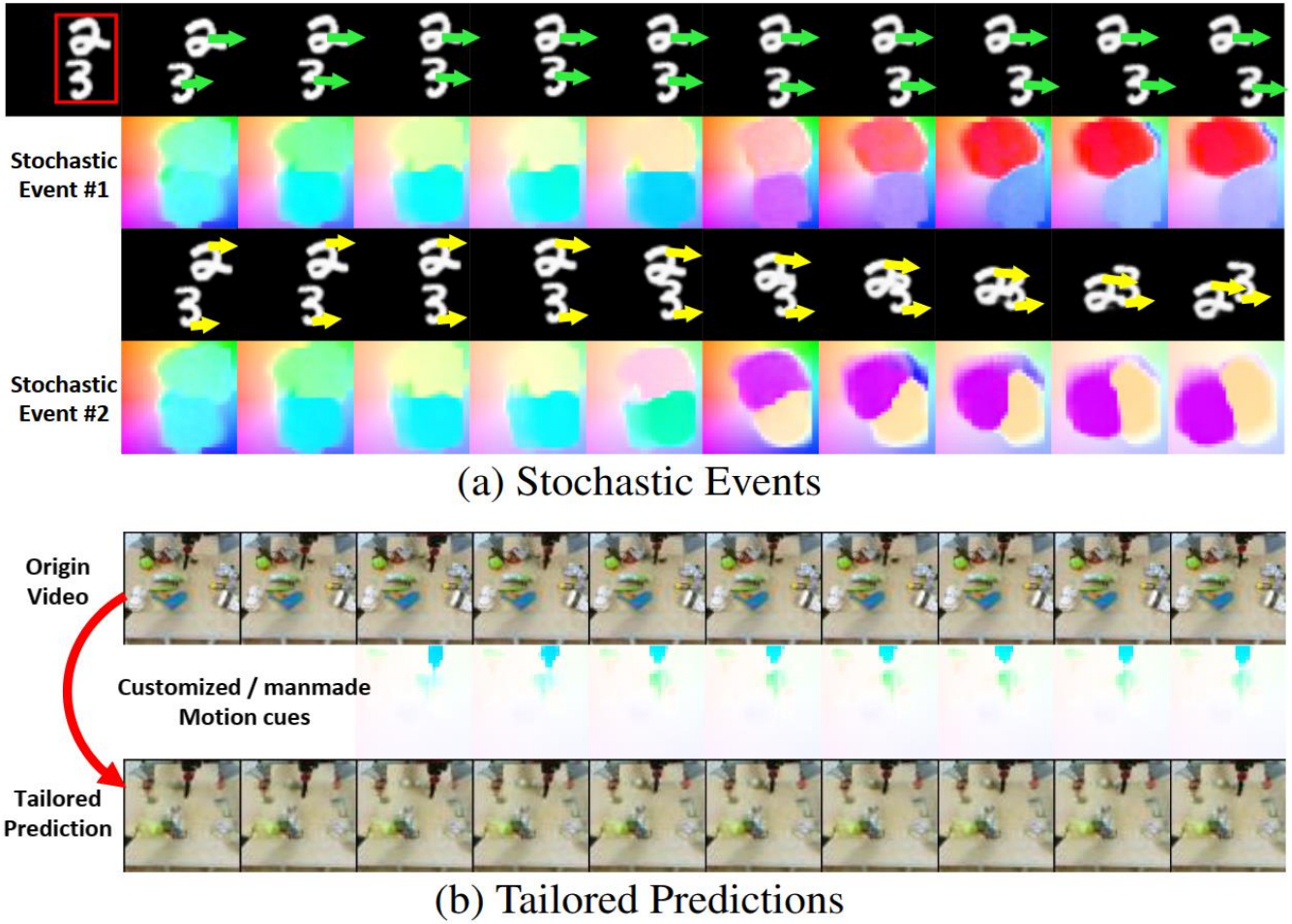
$$f_{x\to g} = \mathrm{softmax}\left(\frac{[\mathcal{T}(f_x)\mathbf{W}^{\mathbf{Q}}][\mathcal{T}(f_g)\mathbf{W}^{\mathbf{K}}]^{\top}}{\sqrt{d}}\right)\mathcal{T}(f_x)\mathbf{W}^{\mathbf{V}}$$

Experimental Results



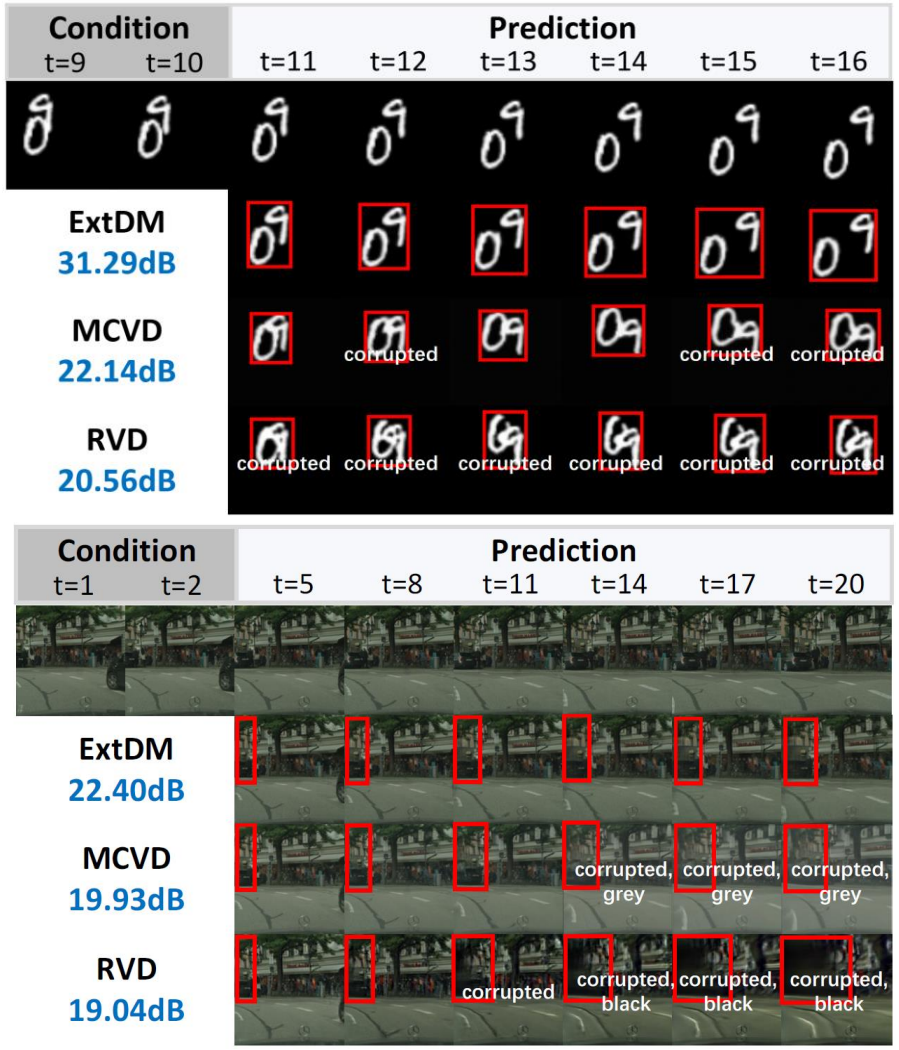
✓ It can predict the videos with **correct trajectories** of objects (**green curve** in the figure).

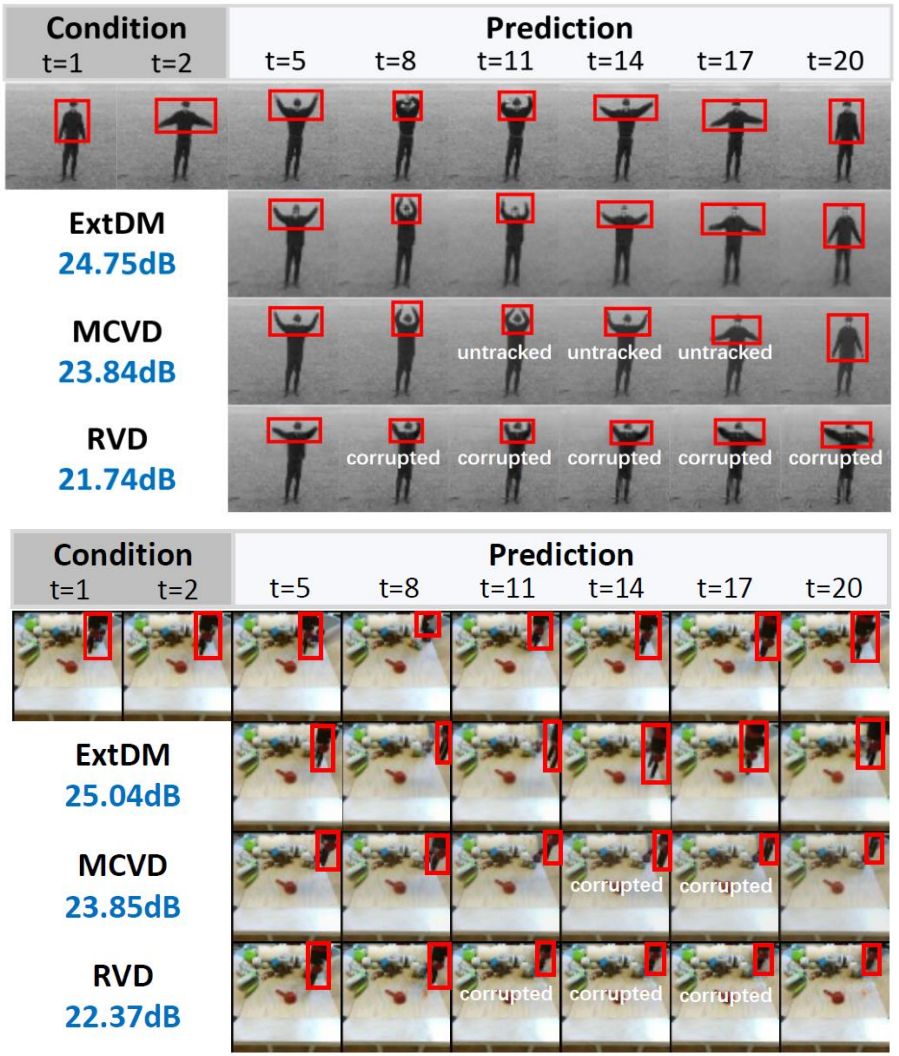
Experimental Results



- ✓ Prediction results can be used to
  (a) generate potential predictions
  (b) customize a preferred trajectory.

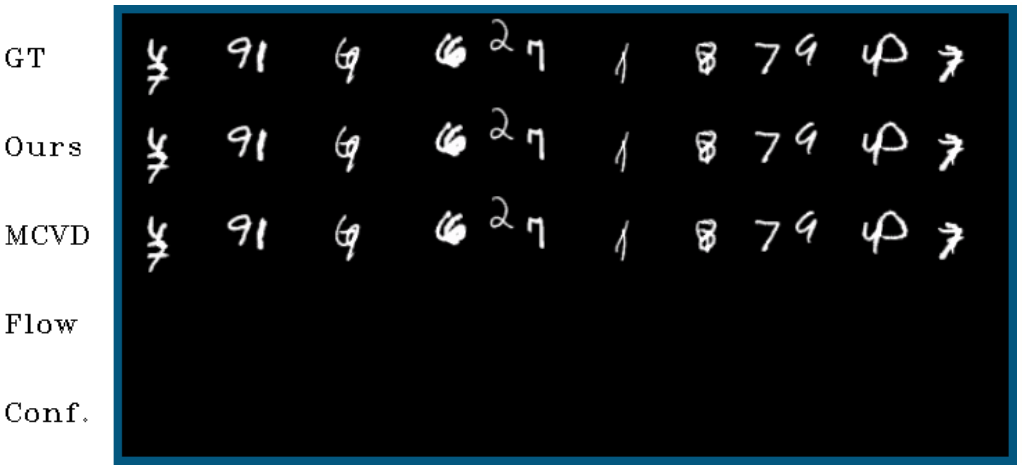
Experimental Results

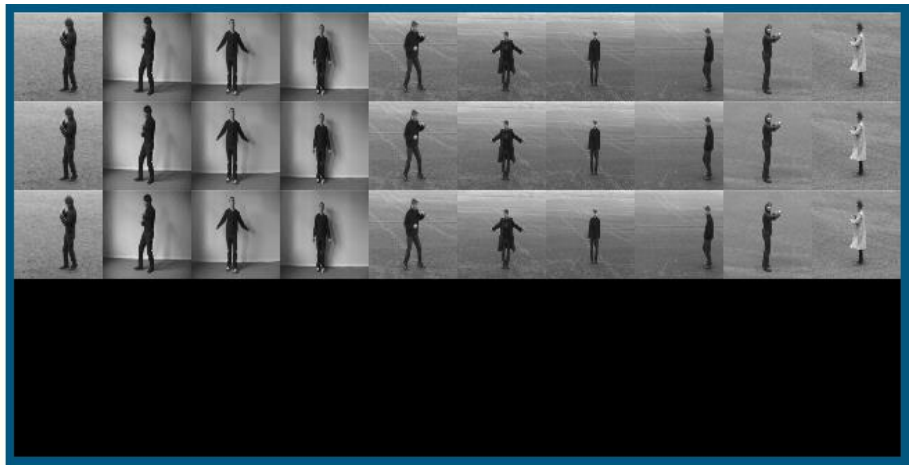


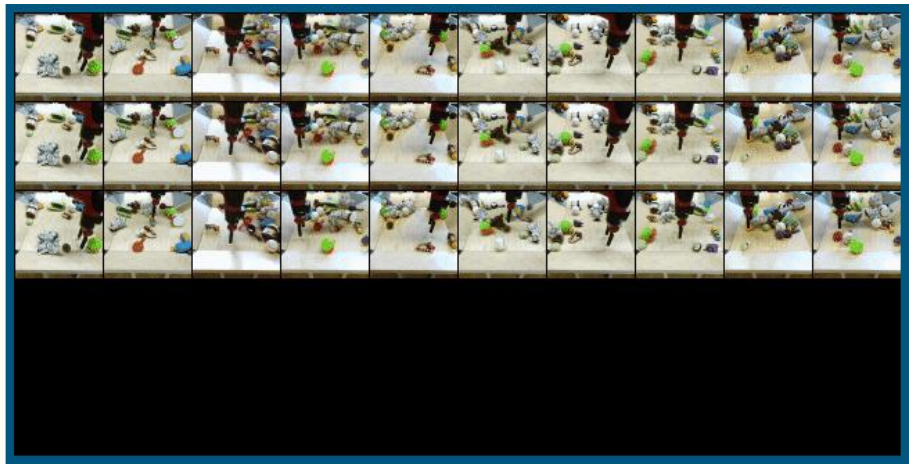


- ✓ **Qualitative comparison on SMMNIST, KTH, Cityscapes and BAIR.**

Experimental Results

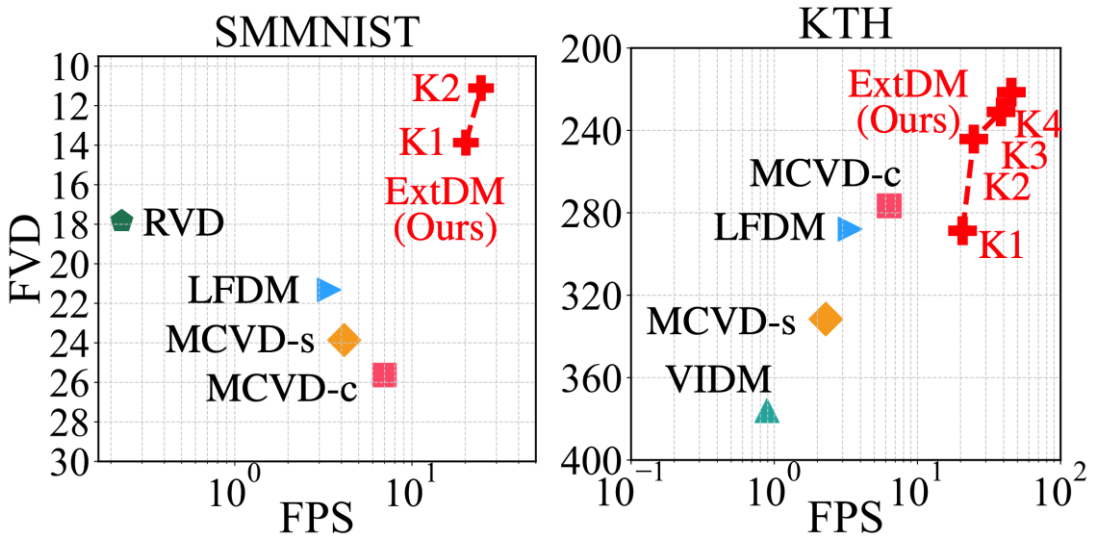


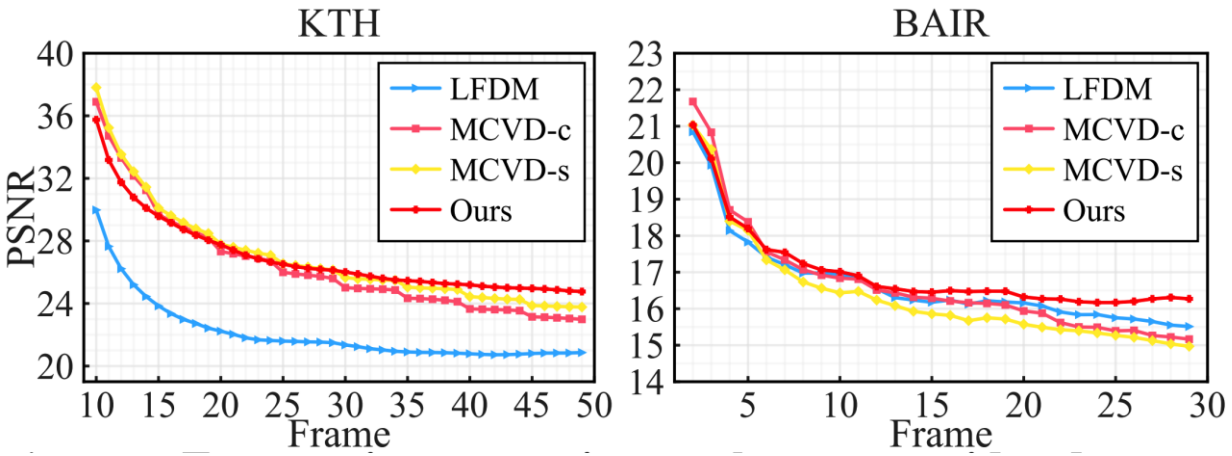




- ✓ **Qualitative comparison on SMMNIST, KTH, Cityscapes and BAIR.**

Experimental Results





- ✓ **Comparison of quality and speed of SOTA DMs for short- and long-term video prediction.**
- ✓ **Frame-wise PSNR comparison on long-term video datasets.**

Thank you

ExtDM: Distribution Extrapolation Diffusion Model for Video Prediction

https://cv.nankai.edu.cn
https://github.com/nku-zhichengzhang/ExtDM

Project Code