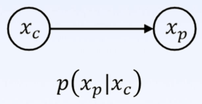
ExtDM: Distribution Extrapolation Diffusion Model for Video Prediction

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

CVPR SEATTLE, WA JUNE 17-21, 2024
VALSE 2024 重庆 视觉与学习青年学者研讨会 VISION AND LEARNING SEMINAR

Introduction

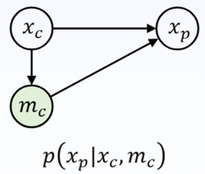
Video prediction aims to **capture the dynamic change** from present x_c to future x_p.



Direct Method

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

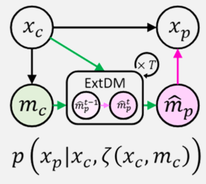
SRVP(ICML20) SimVP(CPVR22)



In-context Learning Method

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

MCNet(ICLR17) MOSO(CVPR23)



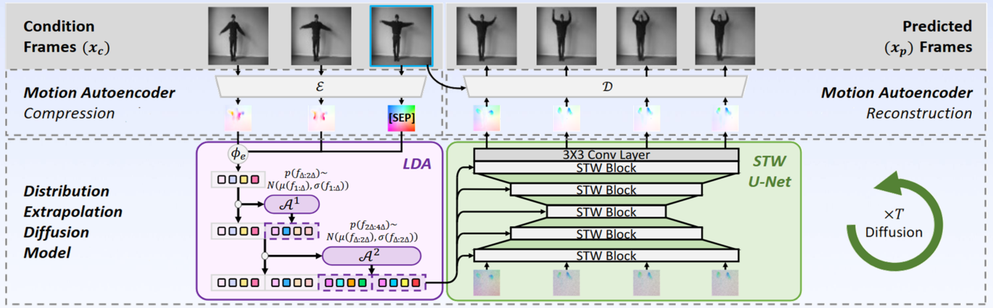
Extrapolation Method (Ours)

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

Contributions

- A **distribution extrapolation diffusion model** that forecasts future frames by extrapolating from the present ones.
- An efficient video prediction 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 **KTH, BAIR, Cityscapes, UCF, and SMMNIST**.

Methodology



Two Mapping Functions

Step 1: $x_c \rightarrow m_c$ & $\hat{m}_p \rightarrow x_p$

- Motion Autoencoder Compression & Reconstruction

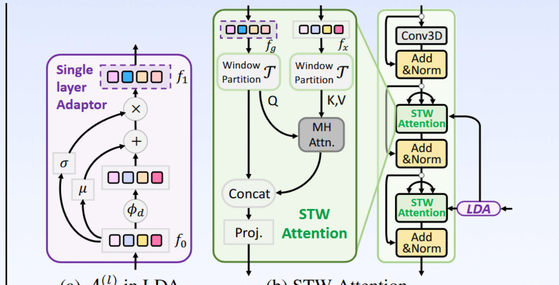
$$m_c = \left\{ m_i \in \mathbb{R}^{3hw} \mid m_i = \mathcal{E}(x_i, x_u) = \begin{bmatrix} w_i \\ o_i \end{bmatrix} \right\}$$

$$x_p = \{ x_j \in \mathbb{R}^{3HW} | x_j = \mathcal{D}(\hat{m}_j, x_u) = \mathcal{G}(o_j \odot \mathcal{W}(z_u, w_j)) \}$$

Step 2: $m_c, x_c \rightarrow \hat{m}_p$

- Distribution Extrapolation Diffusion Model

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



(a) **Layered Distribution Adaptor**

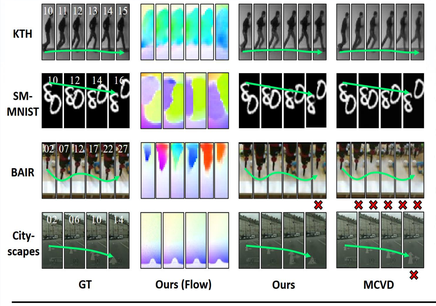
- estimate distribution params
- inference using distribution sampling

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

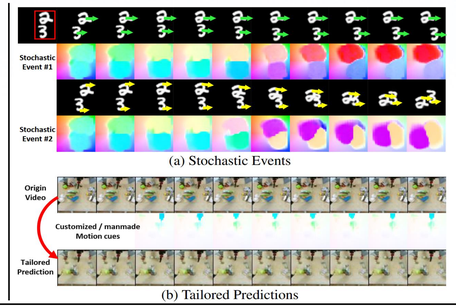
$$f_b = \mathcal{A}(f_a) = (\sigma(f_a) + \sigma')\phi_d\left(\frac{f_a - \mu(f_a)}{\sigma(f_a)}\right) + \mu(f_a) + \mu'$$

(b) **Spatiotemporal Window U-Net**

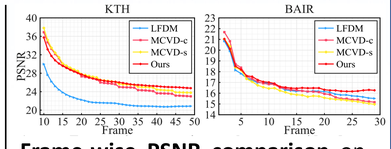
Performance



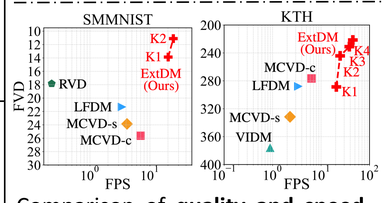
ExtDM can predict the videos with **correct trajectories** of objects (the **green curve** in the left figure).



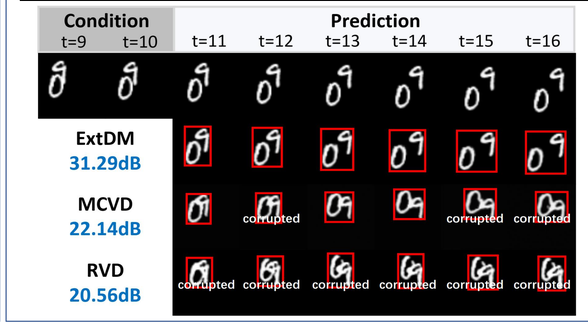
Prediction results can be used to **generate potential predictions** and **customize a preferred trajectory**.

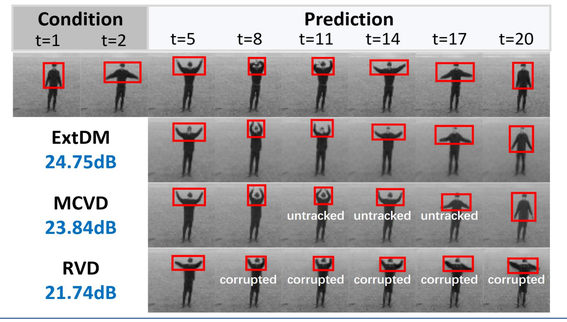


Frame-wise PSNR comparison on long-term video datasets.



Comparison of **quality and speed of SOTA DMs** for short- and long-term video prediction.





Contact

https://zzcheng.top/ExtDM

Project Code