ExtDM: 用于视频预测的分布外推扩散模型*

https://zzcheng.top/ExtDM

张知诚[1,2,†] 胡钧耀[1,2,†] 程文韬[1,‡] Danda Paudel[3,4] 杨巨峰[1,2]

[1] 南开大学计算机学院 VCIP & TMCC & DISSec [2] 南开国际先进研究院（深圳福田）

[3] 苏黎世联邦理工学院计算机视觉实验室 [4] 索菲亚大学 INSAIT 研究所

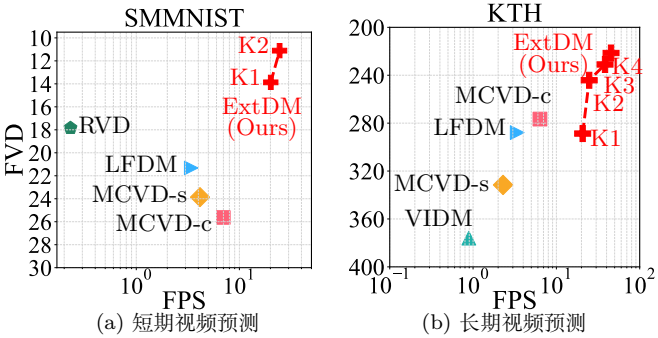
gloryzzc6@sina.com，hujunyao@mail.nankai.edu.cn，danda.paudel@insait.ai，{wentaocheng,yangjufeng}@nankai.edu.cn

摘要

由于未来变化捉摸不定，预测视频后续（*Video Prediction, VP*）充满了挑战，特别是长期视频预测。为建模视频中的时序变化，此前方法利用扩散模型在数据建模方面的能力，使用 *3D* 时空 *U-Net* 对预测的未来帧进行反复修正。然而，这些方法在对齐当前帧与未来帧的差异上仍有不足，并且 *U-Net* 的频繁使用令计算成本显著增加。为此，我们提出一种基于外推扩散模型的视频预测方法 *ExtDM*。该方法通过外推当前帧特征的分布实现预测未来帧。具体来说，我们的方法由三个部分组成：*(i)* 一个运动自编码器，用于在视频帧和运动线索之间进行双射变换；*(ii)* 一个分层分布适配器模块，它在高斯分布的引导下推断当前特征的分布；以及 *(iii)* 一种改进的 *3D U-Net* 架构，其利用时空滑窗注意力机制，专门用于在时间维度上融合引导线索和特征信息。*ExtDM* 在五个流行视频预测基准数据集上进行了广泛实验，其涵盖短长期视频预测，从而验证了该模型的有效性。

1. 引言

视频预测是计算机视觉领域一项长期且具有挑战性的任务。它旨在像人类一样预测视频未来帧，是智能决策系统的关键组成部分 [47]。预测未来事件的像素级可能性，对于系统做出可信决策至关重要，特别是在涉及人类安全的情境中。例如，在自动驾驶汽车穿越拥挤的十字路口时，准确预测行人的移动路径以进行规避是一个重要问题。因此，基于预测的各种下游应用，



(a) 短期视频预测 (b) 长期视频预测

图 1. **本方法与 SOTA 扩散模型的质量和速度对比**。模型分别在 SMMNIST 和 KTH 上进行短期和长期视频预测。我们报告了 FVD 和 FPS 指标。FPS 轴使用对数刻度。

如自动驾驶 [1, 7, 82]、机器人导航 [16, 32, 84]、艺术设计 [19, 41, 55, 56, 70] 以及视频理解 [79, 80, 85, 86]，都正被积极探索并不断应用于生产生活实践。

视频预测的先进工作 [20, 21, 63, 65] 致力于捕捉视频中的动态变化 [2, 18, 35, 59]。直接方法 [5, 26] (图 2 (a)) 仅将 RGB 帧作为输入，但发现视频预测问题因其固有高复杂性而难以直接解决（即估计后验概率 $p(\boldsymbol{x}_p|\boldsymbol{x}_c)$）。因此，基于上下文的学习方法 [54, 60] (图 2 (b)) 将语义线索作为关键信息，无需额外模型进行建模，简单地将运动线索作为隐式引导，即 $p(\boldsymbol{x}_p|\boldsymbol{x}_c, \boldsymbol{m}_c)$。这些方法共同特点是：在短时间内展现出色预测能力，但随时间延长，准确性逐渐下降，有时甚至产生反事实的结果，如视频画面逐渐褪成灰色。这些方法对未来没有构建确定性线索，使得现实与未来仍存在差距。

由于未来充满不确定性，捕捉未来线索的任务极具挑战。这需要了解高层次的时空相关性，并构建能够模拟潜在未来结果的模型。最近的一些方法 [14, 23, 61] 尝试利用流行的视频扩散模型 [25, 49]，并试图重新构思公式来估计未来分布，将这一过程视为串行去噪修

*本文是 CVPR'24 论文 [83] 的中文翻译稿。程文韬为本文的通讯作者。本文由胡钧耀翻译、张知诚校稿。

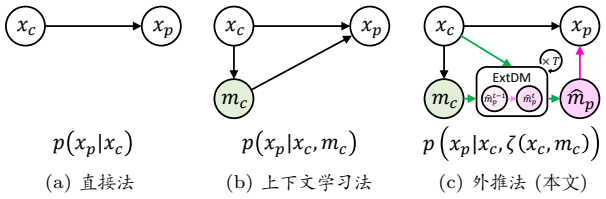


图 2. **ExtDM 与其他视频预测方法的比较**。本图示模型揭示了带有运动线索的预测过程。$\boldsymbol{x}_c, \boldsymbol{x}_p$ 分别表示条件帧和预测帧。$\boldsymbol{m}_c, \hat{\boldsymbol{m}}_p$ 分别表示来自条件帧和本文提出的分布外推扩散模型 ζ 预测的运动线索。$\hat{\boldsymbol{m}}_p^{\{t-1,t\}}$ 是外推运动线索过程的中间结果。

正序列。由于扩散模型具有常微分方程形式，未来帧可以通过一系列链式马尔可夫步骤得到。然而，时空 U-Net 的重复使用导致计算负担显著。这些模型的运行速度往往较慢，其处理速度通常以每秒几帧（FPS）来衡量，如图 1所示。

本工作中，我们提供一种新颖的观点：将视频预测问题视为从现在到未来的确定性运动线索的外推过程，即 $\hat{\boldsymbol{m}}_p = \zeta(\boldsymbol{x}_c, \boldsymbol{m}_c)$，如图 2(c) 所示。我们认为结合未来线索来生成相应帧，要比从零开始生成要容易得多。通过逐渐估计从当前分布偏移的分布，ExtDM 可以保持时间一致性，避免长期视频预测的性能急剧下降。ExtDM 不仅可以产生期望的预测结果，还可以定制未来运动轨迹，并为随机事件生成潜在的建议结果，这些是解耦运动外推信息与实现帧预测的附带产出。此外，运动线索的压缩分辨率降低了预测模型所带来的计算开销。

为了推断未来的运动线索，我们首先用轻量级运动自动编码器从条件帧中提取运动线索。然后，我们提出了一个概率扩散模型，通过一系列马尔可夫步骤外推运动线索。我们利用当前特征来估计视频分布参数。为了解释时序动态，我们提出一种分层分布适配器，其随着时间的推移预测相应的分布参数。这使我们能够根据估计的分布轻松生成未来特征。利用外推得到的预测特征，我们采用稀疏时空窗口 U-Net 融合它与参考特征，以修正预测的未来线索。预测的未来线索最终用于重构视频帧作为外推的视频帧。

我们的贡献有三个方面：(1) 提出了一个分布外推扩散模型，从当前帧外推预测未来帧。(2) 提出了一种有效的视频预测方法，包括视频的压缩和重建。通过模仿未来运动线索，方法可以为随机事件创建量身定制的建议结果。(3) 在五个流行基准上进行大量实验，验证了本方法对短长期视频预测的有效性。

2. 相关工作

视频预测旨在像素级别上实现对未来帧的预测 [4, 9, 10, 22, 69, 74, 78]，并对帧间的变化进行建模 [14, 33, 43, 51, 57, 64, 75]。它对于表征学习 [11, 30, 62, 68]、目标检测 [36–38]、图像分割 [12, 28, 29, 88] 和图像恢复 [42, 91–93] 等下游应用至关重要。早期工作中，[15, 35] 提出了基于随机变分推理的方法，显式地提取空间和时间信息。PRNN [65] 构建时空 LSTM，SLAMP [3] 从外观和光流中学习先验分布，MOSO [54] 将帧映射为运动、场景和对象特征张量。

视频扩散模型将高斯噪声分布转换为视频相关分布 [8, 40, 46, 73, 77]。这个过程依赖于条件引导去噪步骤的多轮迭代。VDM [25] 首次展现了扩散模型应用于视频任务的可行性。RVD [72] 提出了一种预测下一视频帧的残差扩散模型。MCVD [61] 通过压缩维度，实现了一种通用的基于二维卷积的多输入多输出视频扩散模型。RaMViD [27] 引入随机掩码并构造 3D 卷积视频扩散模型。LVDM [24] 使用 3D 自编码器和分层生成机制在潜在空间中生成任何长度的视频。

随着研究不断深入，各种方法都在努力提高未来预测精度 [34, 67, 87, 89, 90]，但由于各方法的固有缺陷，长期动态建模仍具挑战性。直接预测方法带来高计算成本，难以在低资源设备中部署。上下文学习方法从当前帧推断语义线索，与未来帧分布有一定差距。

3. 方法

本文提出的 ExtDM 架构如图 3所示。给定一系列条件帧 $\boldsymbol{x}_c$，ExtDM 需通过充分利用外观和运动线索来预测视频中的未来帧 $\boldsymbol{x}_p$。其中，$\boldsymbol{x}_c, \boldsymbol{x}_p$ 的长度分别为 u, v。本方法的工作流程可概括为三个部分：(i) 运动自编码器压缩部分 (3.1节)，(ii) 分布外推扩散模型 (3.2节)，和 (iii) 运动自动编码器重建部分 (3.1节)。编码器将条件帧 $\boldsymbol{x}_c$ 映射为一系列运动线索 $\boldsymbol{m}_c$（即光流和遮挡图）。接着，分层分布适配器通过一系列高斯过程将特征外推到未来。时空滑窗 U-Net 通过注意力机制将未来特征作为参考信息，从而产生未来的运动线索 $\hat{\boldsymbol{m}}_p$。最终，解码器根据预测出的运动线索 $\hat{\boldsymbol{m}}_p$ 和条件帧 $\boldsymbol{x}_c$ 重建出未来帧 $\boldsymbol{x}_p$。

现在，为了给预测未来帧建立基础，我们建立包括两个映射函数的双射变换：$\boldsymbol{x}_c \to \boldsymbol{m}_c$ 与 $\hat{\boldsymbol{m}}_p \to \boldsymbol{x}_p$。

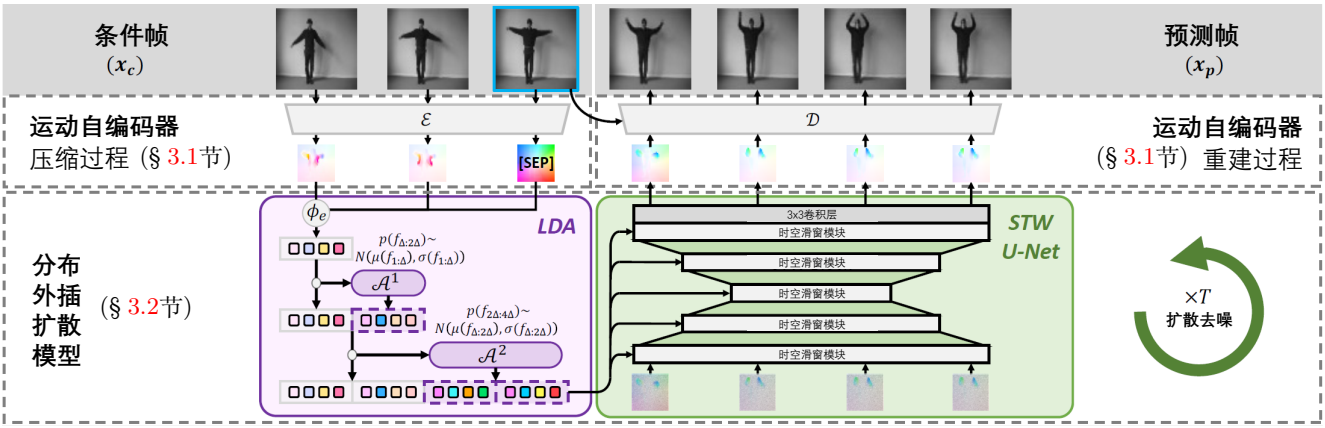


图 3. **ExtDM 架构图示。** ExtDM 由三个主要组件组成：运动自动编码器通过压缩和重构在像素空间和运动空间之间构造双射变换。分层分布适配从条件帧导出的移位分布，通过外推得到未来帧的特征。此外，所构建的时空滑窗 U-Net 以外推特征为指导，在时空维度上进行稀疏跨步注意力学习，以进行特征时空间交互。

3.1. 运动自编码器

如上所述，运动信息为推理未来结果提供了确定性线索。为此，在压缩视频的时候，通过一个轻量级的运动自动编码器完成运动线索和视频帧之间的双射变换。在轻量级自编码器结构基础上 [31]，所构建的运动自编码器由两个阶段组成：编码器 $\mathcal{E}$ 从帧中提取运动线索，解码器 $\mathcal{D}$ 从运动线索重构视频帧。

• 运动自编码器压缩过程 为了从一系列条件帧中提取运动线索，编码器 $\mathcal{E}(\cdot,\cdot)$ 以成对的方式估计视频帧之间的光流和遮挡图。对于长度为 u 的条件帧集 $\boldsymbol{x}_c = \{\boldsymbol{x}_i \in \mathbb{R}^{CHW} \mid i = 1,\dots,u\}$ 我们提取每个条件帧 $\boldsymbol{x}_i$ 与最后一个条件帧 $\boldsymbol{x}_u$（即用于重建的关键帧）之间的运动线索。这些成对的帧被输入到编码器中，以估计它们之间的运动相关性，包括光流 w_i 及其对应的遮挡图 o_i，即：

$$\boldsymbol{m}_c = \left\{\boldsymbol{m}_i \in \mathbb{R}^{3hw} \mid \boldsymbol{m}_i = \mathcal{E}(\boldsymbol{x}_i, \boldsymbol{x}_u) = \begin{bmatrix} w_i \\ o_i \end{bmatrix}\right\}. \quad (1)$$

关键帧的运动线索被一个可学习的标记所替代。为了表征像素偏移，我们计算从条件帧到关键帧的光流 w_i，其形状为 $2 \times h \times w$，用以描述垂直和水平运动。为了模拟具有挑战性的情况，例如被遮挡的背景，我们估计遮挡图 o_i，它指示遮挡的程度，范围从 0 到 1，形状为 $1 \times h \times w$。其中，$h = H/S, w = W/S$，而 S 是降采样因子。

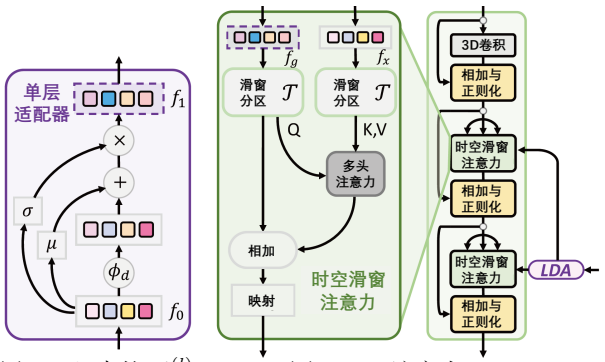
• 运动自编码器重建过程 利用对未来 $\hat{\boldsymbol{m}}_p = \{\hat{\boldsymbol{m}}_j \in \mathbb{R}^{3hw} | j = 1,\dots,v\}$ 的外推运动线索以及关键帧，解码器 $\mathcal{D}(\cdot,\cdot)$ 以类似于编码器的方式，成对地重建未来的帧。我们将关键帧与第 j 个未来帧的预测运动线索看作一组对子。条件帧 $\boldsymbol{z}_u$ 的潜在表示首先在光流 w_j 的指导下进行扭曲。考虑到遮挡，扭曲后的表示通过结合每个预测未来帧的遮挡图 o_j 进一步融合，最终表示为 $o_j \odot \mathcal{W}(\boldsymbol{z}_u, w_j)$。进一步地，将该表示输入网络 $\mathcal{G}$ 以修复被遮挡的区域。在这里，$\mathcal{W}(\boldsymbol{z}, w)$ 是在光流 w 指导下的特征 $\boldsymbol{z}$ 的扭曲操作，而 $\odot$ 是逐元素乘积。最后，得到的重建帧表示为

$$\boldsymbol{x}_p = \{\boldsymbol{x}_j \in \mathbb{R}^{3HW} | \boldsymbol{x}_j = \mathcal{D}(\hat{\boldsymbol{m}}_j, \boldsymbol{x}_u) = \mathcal{G}(o_j \odot \mathcal{W}(\boldsymbol{z}_u, w_j))\}. \quad (2)$$

现在，我们通过估计分布偏移实现外推过程，得到未来运动线索，即 $\boldsymbol{m}_c, \boldsymbol{x}_c \rightarrow \hat{\boldsymbol{m}}_p$。

3.2. 分布外推扩散模型

先进的预测方法通过时间注意力或一维时间卷积网络隐式地利用时间维度的相关性生成未来帧。尽管这些方法在生成未来帧的特征方面是有效的，它们将前一帧编码到网络中以预测未来帧，但这忽略了帧之间的分布先验，并由于未来的不确定性而生成了虚假样本。相比之下，我们提出一个分布外推扩散模型，通过一系列反向（去噪）步骤来外推运动线索 $\hat{\boldsymbol{m}}_p$。基于高斯混合模型的假设，我们设计了一个分层分布适配器，实现对未来特征偏移分布的因果建模，并进一步引入时空注意力来融合外推特征和原始特征。



(a) LDA 中的 $\mathcal{A}^{(l)}$ (b) STW 注意力

图 4. (a) 单层适配器和 (b) 时空窗口块的**详细结构图示**。细节请参阅算法 1。

结合提取的运动线索 $\boldsymbol{m}_c$ 和外观特征的潜在表示 $\boldsymbol{z}_c$，本视频扩散模型包括一个前向函数 $\{q_t\}_{t\in[0,1]}$ 用于向未来帧 $\boldsymbol{m}_p^1 \sim q_1(\boldsymbol{m}_p^1)$ 添加一系列噪声，以及一个反向函数 $\{p_t\}_{t\in[0,1]}$ 通过我们提出的 *时空窗口 U-Net* $\boldsymbol{\epsilon}_\theta(\boldsymbol{m}_p^t, c)$ 从高斯噪声 $p_1(\boldsymbol{m}_p^0) := \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ 预测未来帧。我们将外观特征（潜在表示 $\boldsymbol{z}_c$）和运动特征（运动线索 $\boldsymbol{m}_c$）作为引导 c。为了弥合现在和未来之间的差异，我们通过提出的 *分层分布适配器* 将条件帧 f_c 的引导信息外推到未来帧 f_p。

• 分层分布适配器 由于预测未来的挑战，未来与现在之间存在巨大的差距。为了克服这一挑战，我们将视频帧表示为相同的分布。这使我们能够使用自回归适配器预测未来的样本。我们提出了一种分层分布适配器（LDA），它最初通过“编码”条件帧来估计分布参数，并通过分布采样“推断”未来帧以进行预测。不同于现有的隐式估计未来帧的时间相关性方法，LDA 引入了分布先验作为约束。为更好满足外推需求，LDA 的输入特征的类型包括潜在特征 $\boldsymbol{z}_c$、运动线索 $\boldsymbol{m}_c$ 等。LDA 的伪代码在算法 1中，具体细节在下一段中给出。

给定长度为 Δ 的条件帧，抽取其特征，LDA 产生未来帧的外推特征。输入特征 f_c 首先通过 ϕ_e 进行映射，以利用条件帧之间的时间相关性。然后，各外推层以多层级的方式逐渐外推出未来特征。在第 l 个外推层中，单层适配器将 $\mathcal{A}^{(l)}$ 从当前帧特征 $f_{1:\Delta}$ 外推到未来帧特征 $\widehat{f}_{\Delta:2^l\Delta}$，即：

$$\begin{aligned} f_{1:\Delta} &= \phi_e(f_c), \\ \widehat{f}_{1:2^l\Delta} &= (f_{1:2^{l-1}\Delta}, \mathcal{A}^{(l)}(f_{1:2^{l-1}\Delta})), \\ f_p &= (\widehat{f}_{1:\Delta}, \ldots, \widehat{f}_{2^{L-1}\Delta:2^L\Delta}). \end{aligned} \tag{3}$$

算法 1 LDA: Pytorch 伪代码

```
# f: 输入特征 (N C T H W)
# phi_e: 编码层
# phi_d: 解码层
# L: 层数

f = phi_e(f) # 编码条件帧
for l in range(L):
    r = f
    mu, var = est(f) # 高斯先验
    f_h = (f - mu) / std
    mu = m_est(f_h) + mu
    var = (1 + v_est(f_h)) * var
    f_h = phi_d[l](f_h) # 推断未来帧
    f = f_h * var + mu
    f = torch.cat([r,f], dim=2)

# 分布估计
def est(f, eps=1e-5):
    f_var = f.view(N, C, T, -1).var(dim=3) + eps
    f_std = f_var.sqrt().view(N, C, T)
    f_mean = f.view(N, C, T, -1).mean(dim=3)
    return f_mean, f_std
```

在每一层中，当前帧特征 f_a（例如 $f_{1:\Delta}$）用于近似目标视频分布 $p(vid)$。遵循 [71, 94] 的方法，在 LDA 中，我们将视频的先验分布设定为高斯分布，并得出近似的封闭解，即 $p(vid) \sim \mathcal{N}(\mu(f_a)+\mu', \sigma(f_a)+\sigma')$，其中 μ, σ 分别表示均值和标准差，而 μ', σ' 代表外推更新得到的相应数值。通过这种简化，适配器可以通过算法 1中的 est 函数实现。此外，未来帧特征 f_b 可以从基于推断层 ϕ_d 的特征估计出的分布中进行采样。如图 4 (a) 所示，未来帧特征经推导后可写作：

$$f_b = \mathcal{A}(f_a) = (\sigma(f_a)+\sigma')\phi_d\left(\frac{f_a - \mu(f_a)}{\sigma(f_a)}\right) + \mu(f_a) + \mu'. \tag{4}$$

其中，f_b 与 f_a 的长度相同。

• 时空滑窗 U-Net 遵循 [25] 中的设计，我们引入了一个由 θ 参数化的 3D U-Net $\epsilon_\theta(\boldsymbol{m}_p^t, c)$ 作为去噪器。我们的时空滑窗（STW）U-Net 由各种 STW 块组成，并且具有与普通 U-Net 相同的上采样—下采样结构。在由 LDA 产出的外推特征的引导下，STW U-Net 将带噪声的运动线索作为输入，并迭代地对其进行细化。然而，由于传统 3D 注意力机制计算昂贵，高效地进行引导信息和带噪声特征之间的特征交互是具有挑战性的。为了解决这个问题，本文提出利用时空窗口注意力层来有效地进行它们之间的特征交互。具体而言，使用一个稀疏的交叉注意力机制对特征进行对齐和融合。

为了减少由注意力机制引入的计算开销，STW 注意力在每个滑动窗口中沿着时空维度进行稀疏跨步注

表 1. **KTH 数据集上的消融研究**（$u = 10, v = 40$）。蓝色背景行表示我们方法的最优设置。除非另有说明，否则所有后续实验均使用此设置。CS = 压缩空间。STW = 时空窗口注意力。LDA = 分层分布适配器。OOM = 超出显存。

(a) **模块消融**

CS	STW	LDA	SSIM↑	PSNR↑	LPIPS↓	FVD↓
			0.632	23.43	0.182	422.0
✓			0.749	25.39	0.116	307.9
✓	✓		0.778	26.65	0.109	246.2
✓		✓	0.771	27.12	0.103	243.8
✓	✓	✓	**0.799**	**27.91**	**0.093**	**221.4**

(b) **自编码器重建**（$v = 10$）

数据集	SSIM↑	PSNR↑	LPIPS↓	FVD↓	FPS↑
BAIR	0.951	28.90	0.014	46.1	183.2
Cityscapes	0.858	27.80	0.048	67.0	146.0
KTH	0.911	33.73	0.027	119.7	182.6
SMMNIST	0.986	32.09	0.008	4.8	182.8
UCF	0.890	28.73	0.030	169.9	163.0

(c) **LDA 操作类型**

变量	PSNR↑	FVD↓
Concat	27.22	249.5
AdaIN	26.45	289.4
AdaIN-z	26.83	263.1
Ours	**27.91**	**221.4**

(d) **滑窗大小**

大小	PSNR↑	FVD↓
2	26.98	253.6
4	**27.91**	**221.4**
6	27.24	269.1
8	OOM	OOM

(e) **注意力机制**

类型	PSNR↑	FVD↓
-	27.12	243.8
S-att	27.34	243.7
T-att	27.51	238.9
Ours	**27.91**	**221.4**

(f) **压缩空间类型（PSNR↑）**

压缩空间	KTH	BAIR	City.	SMM.
像素空间 [61]	26.40	17.70	21.90	17.07
潜在空间 [24]	23.43	13.98	14.58	13.38
光流与遮挡图	27.95	18.83	24.34	18.85
光流与遮挡图（真实）	32.11	26.64	23.05	32.14

表 2. **用于 BAIR 数据集的 ExtDM 架构配置**。

模型	ExtDM-K1	ExtDM-K2	ExtDM-K3	ExtDM-K4
LDA 层数	1	2	3	4
预测/条件帧数量比例	1	2	4	8
基础通道数	256	256	256	256
通道倍数	[1,2,4,8]	[1,2,4,8]	[1,2,4,8]	[1,2,4,8]
时空滑窗 [时, 高, 宽]	[2,4,4]	[2,4,4]	[2,4,4]	[2,4,4]
经过压缩的帧分辨率	32	32	32	32

意力学习，该步骤在图 4 (b) 中进行展示。对于来自 LDA 的引导信息 f_g 和待细化的特征 f_x，我们首先将时空特征分割成大小为 k_w 的分区，并在下一个 STW 注意力之前像 [39] 一样移动分区窗口。然后，我们利用跨步和网格滑窗 $\mathcal{T}(\cdot)$ 联合实现时空相干相互作用。因此，我们将交叉注意力构建为如下形式：

$$f_{x\to g} = \text{softmax}(\frac{[\mathcal{T}(f_x)\mathbf{W}^{\mathbf{Q}}][\mathcal{T}(f_g)\mathbf{W}^{\mathbf{K}}]^\top}{\sqrt{d}})\mathcal{T}(f_x)\mathbf{W}^{\mathbf{V}}. \quad (5)$$

其中，$\mathbf{W}^{\mathbf{Q}}, \mathbf{W}^{\mathbf{K}}, \mathbf{W}^{\mathbf{V}}$ 是用于线性映射的可学习矩阵，d 为特征通道数。

基于 STW 注意力机制，构建用于 U-Net 的 STW 模块。STW U-Net 从堆叠的 LDA 中提取特征，通过两个单独的卷积层来估计噪声，其分别负责估计光流和遮挡图。这些噪声被用于预测运动线索 $\hat{\boldsymbol{m}}_p$。

4. 实验

设置。本文遵循 [23, 61] 的实验设置，在五个数据集上进行长短视频预测实验，包括 KTH [50], BAIR [17], Cityscapes [13], SMMNIST [15, 53] 和 UCF-101 [52]。ExtDM 训练需经过两个阶段：(a) 优化自编码器感知损失 [31] 和 (b) 优化扩散模型 L2 损失 [8]。不同 ExtDM 对应于不同数量的 LDA 层数，如表 2所示。

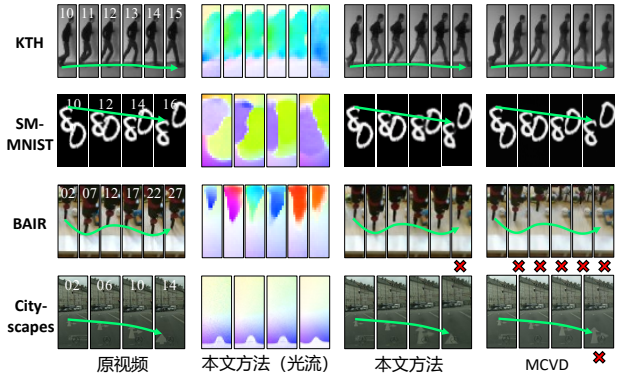


图 5. **本方法与 SOTA 方法的定性比较**。画面中每个重要目标的轨迹由绿色曲线表示。

评估指标。遵循 [61, 76] 的评估方案，本文采用峰值信噪比（PSNR）、结构相似性指数（SSIM）[66]、感知图像质量评估（LPIPS）[81] 和弗雷歇距离（FVD）[58] 来评估生成视频的质量。此外，本文通过报告模型每秒可预测的帧数（FPS）以评估各方法的推理效率。

4.1. 消融实验

为研究运动自编码器、分层分布适配器和时空滑注意力的有效性，本文在表 1(a) 中进行模块消融实验。我们通过深入分析各个组件的不同变体来进一步讨论每个组件，以回答以下研究问题。**问题 1:** 哪个预测空间表现更好？**问题 2:** 如何将特征外推到未来？**问题 3:** 在 STW 不同设置中，哪种大小的窗口更好？**问题 4:** 哪种注意力机制可以更有效地融合外推特征？

问题 1: 近期，潜在空间因其将视频作为紧凑的低维表示来降低计算成本，这引起了研究者的兴趣。在短期和长期视频预测数据集上进行实验，如表 1 (b) 和 (f)

表 3. **KTH 数据集（64 × 64）上的定量比较**。我们在两种设置下与 10 种 SOTA 方法进行比较。**粗体**格式和下划线 格式分别表示最高和第二高的性能。K 代表 LDA 的层数。

方法	年份	$u=10, v=30$				$u=10, v=40$				FPS↑
		SSIM↑	PSNR↑	LPIPS↓	FVD↓	SSIM↑	PSNR↑	LPIPS↓	FVD↓	
U-ViT[6]	CVPR23	0.642	26.13	0.155	694.7	0.606	25.40	0.179	772.0	4.08
DiT[48]	ICCV23	0.657	24.29	0.124	750.4	0.641	23.56	0.145	712.8	2.22
RaMViD [27]	TMLR24	0.590	23.42	0.169	581.3	0.567	22.96	0.187	571.6	0.12
LVDM [24]	ArXiv23	0.644	23.83	0.167	481.1	0.632	23.43	0.182	422.0	1.77
RVD [72]	ArXiv22	0.782	25.32	0.128	441.1	0.758	24.45	0.152	419.1	0.23
VIDM [44]	AAAI23	0.694	25.02	0.150	357.1	0.661	24.32	0.172	376.0	0.89
LFDM [45]	CVPR23	0.772	26.89	0.110	320.2	0.750	26.41	0.116	287.9	3.39
MCVD-c [61]	NeurIPS22	0.812	27.45	0.108	299.8	0.793	26.20	0.124	276.6	6.35
MCVD-cpf [61]	NeurIPS22	0.746	24.30	0.143	294.9	0.720	23.48	0.173	368.4	6.38
MCVD-s [61]	NeurIPS22	0.835	27.50	0.092	323.0	0.744	26.40	0.115	331.6	2.29
ExtDM-K1	CVPR24	0.804	28.34	0.077	284.9	0.784	27.89	0.090	288.7	20.67
ExtDM-K2	CVPR24	0.801	28.43	0.076	239.8	0.779	27.73	0.089	244.1	24.76
ExtDM-K3	CVPR24	0.817	**29.04**	**0.071**	238.3	0.787	**28.31**	**0.084**	231.0	38.36
ExtDM-K4	CVPR24	**0.838**	28.53	0.082	**227.9**	**0.799**	27.91	0.093	**221.4**	**45.28**

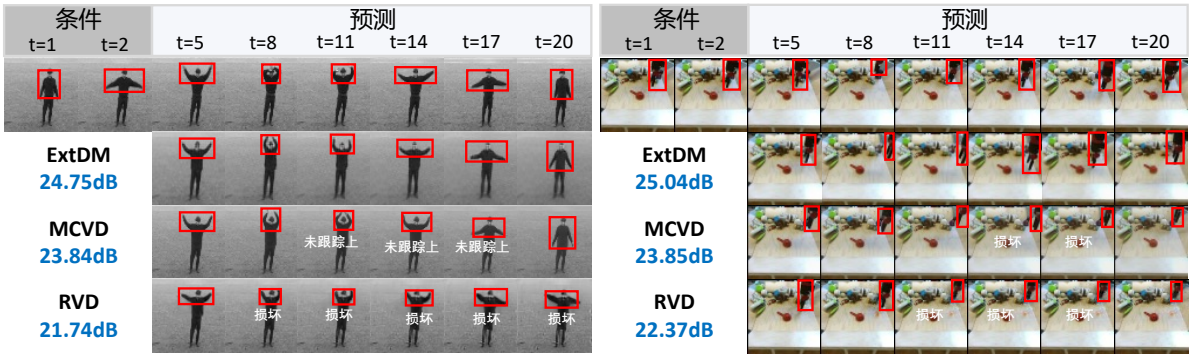


图 6. **KTH Action（左）和 BAIR（右）的定性比较**。重点关注目标（人类手臂和机器人手臂）用红色框标示。相应 PSNR 数值展示在每个视频下方。更多结果请参见表 3和表 4。

所示。可见，该空间与其他预测空间相比具有较高的 PSNR，展现出色的重建质量。这种改进归因于运动线索捕获的时间一致性，其中，ExtDM 有效地将外观信息与运动线索一起融入预测中。此外，可以通过结合来自真实数据的精确光流和遮挡图（灰色背景行）来进一步提高重建质量。

问题 2: 弥合未来与现在之间的差距是视频预测的关键点。本文进行实验以验证 LDA 的有效性，在表 1 (a) 和 (c) 进一步讨论其变体。由其可知，适配器在消融实验中获得了 9.24% 的性能提升，并且以 6.89% 优势超越第二名。这种改进是由于 LDA 通过确定性轨迹外推，帮助模型避免了不确定性，最终生成了可信的预测结果。如图 5所示，我们的方法可以预测具有正确轨迹的视频后续。相比之下，其他方法直接基于外观预测视频，在视频结尾会产生模糊的帧。

问题 3: 随着时空窗口变小，引导特征与普通特征之间的特征交互愈加高效。因此，本文研究了窗口大小效应，如表 1 (a) 和 (d) 中所示。结果表明，将窗口大小设置为 4 时，本方法成功在有效性和效率之间取得平衡。原因有两方面。首先，适当的大小能够控制算力成本，同时在过滤不相关特征方面仍然实现相当的性能，特别是在视频物体具有遮挡关系的情况下。另一方面，较大的窗口大小可能会导致过度计算，并且可能无法最优地捕获相关信息，从而导致性能不佳。

问题 4: 融合外推特征可极大促进当前与未来间的特征交互。我们在表 1 (a) 和 (e) 中实验，以验证 STW 注意力的有效性。分析可知，特征融合不仅需要在时间维度，还需要在空间维度上进行，以充分利用指导信息。

4.2. 对比实验

如表 3，表 4，表 5，表 6与表 7中所示，我们在两个短期视频数据集（即 SMMNIST 和 UCF-101）和三个长期视频数据集（即 KTH、BAIR 和 Cityscapes）上展示了主要结果。

表 4. **BAIR 数据集上的定量比较（尺寸 64×64）**。我们在两个预测时长下比较我们的方法与 12 个 SOTA 方法的结果。

方法	年份	$u=2, v=14$				$u=2, v=28$				FPS↑
		SSIM↑	PSNR↑	LPIPS↓	FVD↓	SSIM↑	PSNR↑	LPIPS↓	FVD↓	
DiT[48]	ICCV23	0.543	15.07	0.171	1013.6	0.520	14.65	0.186	2290.7	2.25
LVDM [24]	ArXiv23	0.464	14.40	0.182	900.6	0.435	13.98	0.198	1663.8	1.31
U-ViT[6]	CVPR23	0.740	17.32	0.078	200.2	0.696	16.53	0.094	263.5	3.84
LFDM [45]	CVPR23	0.770	17.45	0.084	167.6	0.730	16.68	0.106	276.8	5.66
RVD [72]	ArXiv22	0.792	17.88	0.072	139.7	0.750	16.76	0.093	267.1	0.23
RaMViD [27]	TMLR24	0.758	17.55	0.085	166.5	0.691	16.51	0.109	238.7	0.41
VIDM [44]	AAAI23	0.763	16.97	0.080	131.7	0.728	16.20	0.096	194.6	0.82
MCVD-c [61]	NeurIPS22	0.834	19.10	0.078	90.5	0.785	17.60	0.100	120.6	4.15
MCVD-cp [61]	NeurIPS22	0.838	19.10	0.075	87.8	0.797	17.70	0.078	119.0	4.15
MCVD-cpf [61]	NeurIPS22	0.787	17.10	0.077	89.6	0.745	16.20	0.086	118.4	4.15
MCVD-s [61]	NeurIPS22	0.836	19.10	0.078	94.1	0.779	17.50	0.108	132.1	2.51
MCVD-sp [61]	NeurIPS22	0.837	19.20	0.076	90.5	0.789	17.70	0.097	127.9	2.51
ExtDM-K1	CVPR24	0.785	17.73	0.077	114.2	0.748	17.04	0.096	140.3	29.32
ExtDM-K2	CVPR24	0.827	19.76	0.078	97.1	0.790	18.53	0.073	125.8	35.31
ExtDM-K3	CVPR24	0.838	**20.18**	0.066	86.1	0.802	**18.83**	**0.069**	114.7	**37.44**
ExtDM-K4	CVPR24	**0.845**	20.04	**0.053**	**81.6**	**0.814**	18.74	**0.069**	**102.8**	**47.01**

表 5. **Cityscapes 数据集上的定量比较（尺寸 128×128）**。

方法	年份	$u=2, v=28$				FPS↑
		SSIM↑	PSNR↑	LPIPS↓	FVD↓	
U-ViT[6]	CVPR23	0.362	10.84	0.431	1045.3	0.40
RaMViD [27]	TMLR24	0.454	13.14	0.395	812.6	0.12
VIDM [44]	AAAI23	0.539	18.49	0.252	724.7	0.54
RVD [72]	ArXiv22	0.489	17.21	0.242	465.0	0.15
LFDM [45]	CVPR23	0.579	20.32	0.157	194.9	2.93
MCVD-c [61]	NeurIPS22	0.690	21.90	0.112	141.3	2.26
MCVD-s [61]	NeurIPS22	0.720	22.50	0.121	184.8	0.89
ExtDM-K1	CVPR24	0.631	21.49	0.145	157.2	24.65
ExtDM-K2	CVPR24	0.683	21.72	0.135	152.8	28.60
ExtDM-K3	CVPR24	0.701	22.42	0.126	137.2	30.46
ExtDM-K4	CVPR24	**0.745**	**22.84**	**0.108**	**121.3**	**35.44**

表 6. **SMMNIST 数据集上的定量比较（尺寸 64×64）**。

方法	年份	$u=10, v=10$				FPS↑
		SSIM↑	PSNR↑	LPIPS↓	FVD↓	
U-ViT[6]	CVPR23	0.510	17.44	0.138	251.5	4.08
RaMViD [27]	TMLR24	0.585	18.30	0.123	100.4	0.12
LVDM [24]	ArXiv23	0.624	13.38	0.198	49.85	1.77
VIDM [44]	AAAI23	0.514	12.06	0.241	49.08	0.86
LFDM [45]	CVPR23	0.710	15.68	0.137	21.32	3.41
MCVD-c [61]	NeurIPS22	0.786	17.22	0.117	25.63	6.99
MCVD-cpf [61]	NeurIPS22	0.753	16.33	0.139	20.77	6.92
MCVD-s [61]	NeurIPS22	0.785	17.07	0.129	23.86	4.15
MCVD-sf [61]	NeurIPS22	0.758	16.31	0.141	44.14	4.09
MCVD-spf [61]	NeurIPS22	0.748	16.15	0.146	36.12	4.10
RVD [72]	ArXiv22	0.764	18.56	0.123	17.84	0.23
ExtDM-K1	CVPR24	0.776	17.55	0.085	13.87	20.16
ExtDM-K4	CVPR24	**0.813**	**19.59**	**0.068**	**11.11**	**24.54**

表 7. **UCF-101 数据集上的定量比较（尺寸 64×64）**。

方法	年份	$u=4, v=12$				FPS↑
		SSIM↑	PSNR↑	LPIPS↓	FVD↓	
RaMViD [27]	TMLR24	0.639	21.37	0.090	396.7	0.33
LFDM [45]	CVPR23	0.627	20.92	0.098	698.2	3.53
MCVD-cp [61]	NeurIPS22	0.658	21.82	0.088	468.1	1.72
ExtDM-K2	CVPR24	**0.754**	**23.89**	**0.056**	**394.1**	**39.80**

短期视频预测。所提出模型比替代方案 (RVD) 提升 23.60%。将视频的运动和外观解耦到自编码器以及扩散模型中，使我们能够识别关键的运动和内容特征，并产生与其他方法相比具有高保真度的预测。本方法避免了产生事实性错误样本，如随时间变化的数字。

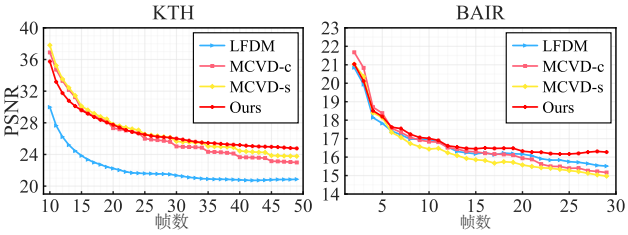


图 7. **长期视频数据集上的逐帧比较结果**。

长期视频预测。首先，可以观察到在二十个指标上，三个基准测试的性能提升了 10.23%，这验证了 ExtDM 与先进 SOTAs（例如 MCVD、RVD）中的次优指标相比的有效性。其次，在 KTH 和 BAIR 两种预测设置下，发现预测长视频的性能下降（与预测短视频的设置相比）为-9.12%，优于高级 SOTA 方法（MCVD）的-14.60%。第三，我们注意到 LDA 的使用可以显著提高性能。与单层 LDA 设置（K1）相比，多层 LDA 设置在 KTH、BAIR、Cityscapes 和 SMMNIST 上的平均性能分别提高了 4.73%, 19.41%, 18.17% 与 14.07%。这是因为时间分布很难在一段时间内估计，我们将其解耦为多个外推步骤可以产生更好的结果。此外，由于引入多层增加了计算成本，导致处理速度稍慢。因此，ExtDM 提供了可扩展的变体，无论是需要精确的性能还是仅需高速执行，不同场景的要求都可以满足。

逐帧效果比较。为了更好地利用时序一致性，我们绘制了测试视频的平均 PSNR—视频帧索引的折线图。如图 7所示，我们计算第一帧和最后一帧之间的性能下降情况。ExtDM 在长期视频预测中展现低退化性能，实现比 MCVD (-7.87 与 -10.20) 好 29.60% 的性能。

可视化效果比较。图 6 与 图 8 显示了 ExtDM 和 SOTA 方法之间的定性比较。我们发现，ExtDM 可以产生时

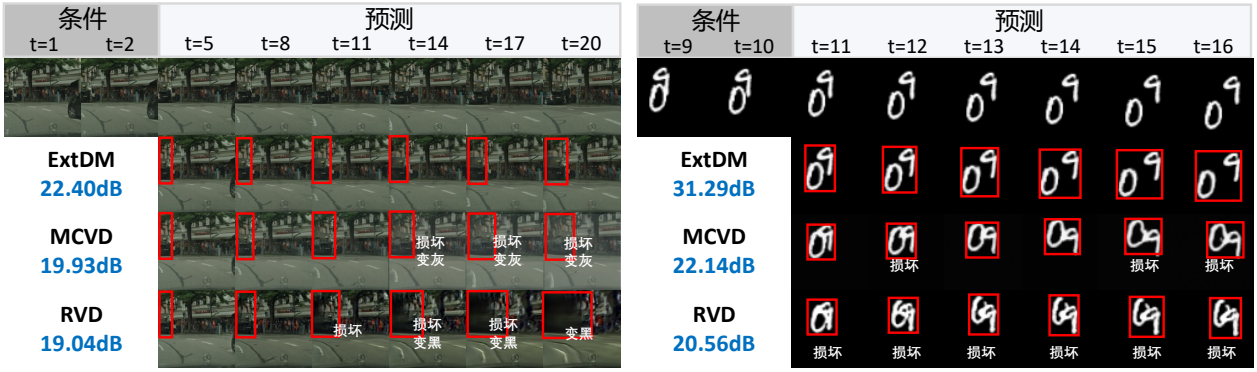


图 8. **Cityscapes 和 Stochastic Moving MNIST 数据集上的定性比较**。目标（左边的汽车和右边的两个数字）由红框表示。对应的 PSNR 在每个视频下方示出。有关更多结果，请参阅表 5 和表 6。

序一致的结果，并避免在现有 SOTA 方法中出现的事实错误（例如，不正确的运动，逐渐灰色，损坏的人，数字和背景）。这再次验证了我们的洞察，即现有方法主要关注来自当前的线索，而忽视了未来的动态变化。

运行时间分析。我们在五个数据集上的实验结果验证了方法的效率，模型运行速度提高了 7.51 倍。为避免计算高分辨率帧所需要的沉重计算成本，本方法估计的是低分辨率的运动线索。

5. 讨论

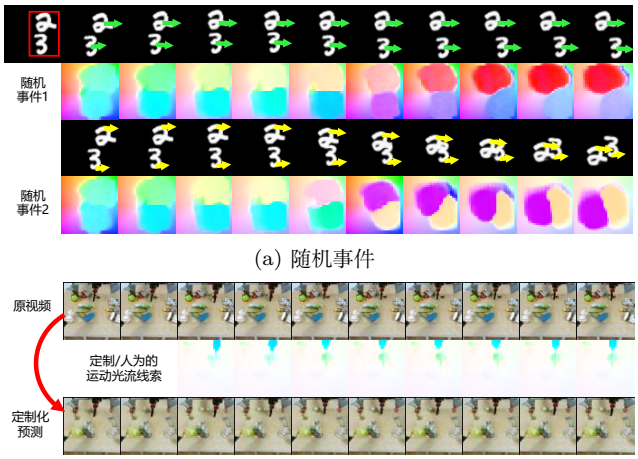
基于 ExtDM 的视频生成框架可以推进多个研究方向。在此，我们设想了两个潜在应用前景。

随机事件。长久以来，对随机事件进行预测是一个理想的追求。借助 ExtDM 的支持，现在能够朝这一目标迈出更坚实的步伐。如果随机事件遵循物理规律，例如到达边界时改变方向和速度 [15]，我们可以相应地设计一些规则来生成运动线索。ExtDM 可根据规则性的运动线索自然地生成潜在预测结果，如图 9 (a) 所示。

定制化预测。除生成随机事件外，用户也可以定制偏好的运动轨迹。基于 ExtDM，可将视频预测解耦为运动线索外推和视频帧重建结果两个部分，从而使预测结果能够根据人为的运动线索进行修正。图 9 (b) 展示了通过从另一视频中提取运动线索进行定制化预测。

6. 总结

本文提出了一种基于扩散模型的框架 ExtDM，其通过沿时间维度外推分布进行视频预测。我们将视频预测任务重新定义为估计视频中的未来线索，从而提供了一个估计当前到未来外推分布偏移量的解决方案。该外推过程主要关注：i) 对短期和长期未来的时序动态进行建模，以及 ii) 以成对的方式沿时间维度构建的视频分布。作为一个容易实现的目标，外推线索由于其经过压缩过程，可被被高效预测，并且可以通过人为定制轨迹，以适应未来多种可能情况。总体而言，ExtDM 使视频预测更加有效，并且可从修正未来线索的角度进行解释。在五个视频预测数据集上的广泛实验表明，本模型在短期和长期预测方面都达到新的 SOTA。



(a) 随机事件

(b) 定制化预测

图 9. (a) SMMNIST 上的随机事件和 (b) BAIR 上的定制化的预测结果。

致谢 本论文得到了天津市自然科学基金（No.20JCJQJC00020）、国家自然科学基金（No.62306154）、中央高校基础研究基金、南开大学超算中心（NKSC）的资助，以及保加利亚教育和科学部对 INSAIT 的部分支持。

References

[1] M. Abraham, N. Suryawanshi, N. Joseph, and D. Hadsul. Future predicting intelligent camera security system. In *ICIT*, 2021. 1

[2] D. Acharya, Z. Huang, D. P. Paudel, and L. Van Gool. Towards high resolution video generation with progressive growing of sliced wasserstein gans. *arXiv*, 2018. 1

[3] A. K. Akan, E. Erdem, A. Erdem, and F. Güney. Slamp: Stochastic latent appearance and motion prediction. In *ICCV*, 2021. 2

[4] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine. Stochastic variational video prediction. In *ICLR*, 2017. 2

[5] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine. Stochastic variational video prediction. In *ICLR*, 2018. 1

[6] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, and J. Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, 2023. 6, 7

[7] A. Bhattacharyya, M. Fritz, and B. Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *CVPR*, 2018. 1

[8] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 2, 5

[9] L. Castrejon, N. Ballas, and A. Courville. Improved conditional vrnns for video prediction. In *ICCV*, 2019. 2

[10] Z. Chang, X. Zhang, S. Wang, S. Ma, and W. Gao. Strpm: A spatiotemporal residual predictive model for high-resolution video prediction. In *CVPR*, 2022. 2

[11] X. Chen, L. Fang, L. Ye, and Q. Zhang. Deep video harmonization by improving spatial-temporal consistency. *MIR*, 2024. 2

[12] Y. Chen, C. Hao, Z.-X. Yang, and E. Wu. Fast target-aware learning for few-shot video object segmentation. *SCIS*, 2022. 2

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 5

[14] A. Davtyan, S. Sameni, and P. Favaro. Randomized conditional flow matching for video prediction. *arXiv*, 2022. 1, 2

[15] E. Denton and R. Fergus. Stochastic video generation with a learned prior. In *ICML*, 2018. 2, 5, 8

[16] F. Ebert, C. Finn, S. Dasari, A. Xie, A. Lee, and S. Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv*, 2018. 1

[17] F. Ebert, C. Finn, A. X. Lee, and S. Levine. Self-supervised visual planning with temporal skip connections. In *CoRL*, 2017. 5

[18] J.-Y. Franceschi, E. Delasalles, M. Chen, S. Lamprier, and P. Gallinari. Stochastic latent residual video prediction. In *ICML*, 2020. 1

[19] T.-J. Fu, X. E. Wang, S. T. Grafton, M. P. Eckstein, and W. Y. Wang. M3l: Language-based video editing via multi-modal multi-level transformers. In *CVPR*, 2022. 1

[20] A. Furnari and G. M. Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *ICCV*, 2019. 1

[21] Z. Gao, C. Tan, L. Wu, and S. Z. Li. Simvp: Simpler yet better video prediction. In *CVPR*, 2022. 1

[22] V. L. Guen and N. Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *CVPR*, 2020. 2

[23] W. Harvey, S. Naderiparizi, V. Masrani, C. Weilbach, and F. Wood. Flexible diffusion modeling of long videos. In *NeurIPS*, 2022. 1, 5

[24] Y. He, T. Yang, Y. Zhang, Y. Shan, and Q. Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv*, 2023. 2, 5, 6, 7

[25] J. Ho, T. Salimans, A. A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. In *NeurIPS*, 2022. 1, 2, 4

[26] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *JMLR*, 2013. 1

[27] T. Höppe, A. Mehrjou, S. Bauer, D. Nielsen, and A. Dittadi. Diffusion models for video prediction and infilling. *TMLR*, 2024. 2, 6, 7

[28] D. Huang, X. Xiong, D.-J. Fan, F. Gao, X.-J. Wu, and G. Li. Annotation-efficient polyp segmentation via active learning. *arXiv*, 2024. 2

[29] D. Huang, X. Xiong, J. Ma, J. Li, Z. Jie, L. Ma, and G. Li. Alignsam: Aligning segment anything model

to open context via reinforcement learning. In *CVPR*, 2024. 2

[30] G. Jia and J. Yang. S 2-ver: Semi-supervised visual emotion recognition. In *ECCV*, 2022. 2

[31] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 3, 5

[32] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *TPAMI*, 2015. 1

[33] Y.-H. Kwon and M.-G. Park. Predicting future frames using retrospective cycle gan. In *CVPR*, 2019. 2

[34] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 2015. 2

[35] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine. Stochastic adversarial video prediction. *arXiv*, 2018. 1, 2

[36] J. Li, J. Zhang, J. Li, G. Li, S. Liu, L. Lin, and G. Li. Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *CVPR*, 2024. 2

[37] X. Liu, G. Xiao, R. Chen, and J. Ma. Pgfnet: Preference-guided filtering network for two-view correspondence learning. *TIP*, 2023.

[38] X. Liu and J. Yang. Progressive neighbor consistency mining for correspondence pruning. In *CVPR*, 2023. 2

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 5

[40] H. Lu, G. Yang, N. Fei, Y. Huo, Z. Lu, P. Luo, and M. Ding. Vdt: An empirical study on video diffusion with transformers. *arXiv*, 2023. 2

[41] S. Ma, J. Gao, R. Wang, J. Chang, Q. Mao, Z. Huang, and C. Jia. Overview of intelligent video coding: from model-based to learning-based approaches. *VI*, 2023. 1

[42] S. Ma, L. Zhang, S. Wang, C. Jia, S. Wang, T. Huang, F. Wu, and W. Gao. Evolution of avs video coding standards: twenty years of innovation and development. *SCIS*, 2022. 2

[43] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2017. 2

[44] K. Mei and V. M. Patel. Vidm: Video implicit diffusion models. *AAAI*, 2023. 6, 7

[45] H. Ni, C. Shi, K. Li, S. X. Huang, and M. R. Min. Conditional image-to-video generation with latent flow diffusion models. In *CVPR*, 2023. 6, 7

[46] Y. Nikankin, N. Haim, and M. Irani. Sinfusion: Training diffusion models on a single image or video. In *ICML*, 2023. 2

[47] S. Oprea, P. Martinez-Gonzalez, A. Garcia-Garcia, J. A. Castro-Vargas, S. Orts-Escolano, J. Garcia-Rodriguez, and A. Argyros. A review on deep learning techniques for video prediction. *TPAMI*, 2020. 1

[48] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 6, 7

[49] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[50] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, 2004. 5

[51] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, 2015. 2

[52] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv*, 2012. 5

[53] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015. 5

[54] M. Sun, W. Wang, X. Zhu, and J. Liu. Moso: Decomposing motion, scene and object for video prediction. *arXiv*, 2023. 1, 2

[55] B. Tan, N. Xue, S. Bai, T. Wu, and G.-S. Xia. Planetr: Structure-guided transformers for 3d plane recovery. In *ICCV*, 2021. 1

[56] B. Tan, N. Xue, T. Wu, and G.-S. Xia. Nope-sac: Neural one-plane ransac for sparse-view planar 3d reconstruction. *TPAMI*, 2024. 1

[57] S. Tian, C. Finn, and J. Wu. A control-centric benchmark for video prediction. In *ICLR*, 2023. 2

[58] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv*, 2018. 5

[59] A. Villar-Corrales, A. Karapetyan, A. Boltres, and S. Behnke. Mspred: Video prediction at multiple spatio-temporal scales with hierarchical recurrent net-

works. In *BMVC*, 2022. 1

[60] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing motion and content for natural video sequence prediction. In *ICLR*, 2017. 1

[61] V. Voleti, A. Jolicoeur-Martineau, and C. Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. In *NeurIPS*, 2022. 1, 2, 5, 6, 7

[62] L. Wang, G. Jia, N. Jiang, H. Wu, and J. Yang. Ease: Robust facial expression recognition via emotion ambiguity-sensitive cooperative networks. In *ACM MM*, 2022. 2

[63] Y. Wang, Z. Gao, M. Long, J. Wang, and P. S. Yu. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *ICML*, 2018. 1

[64] Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei. Eidetic 3d lstm: A model for video prediction and beyond. In *ICLR*, 2018. 2

[65] Y. Wang, H. Wu, J. Zhang, Z. Gao, J. Wang, P. Yu, and M. Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *TPAMI*, 2022. 1, 2

[66] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 5

[67] C. Wen, G. Jia, and J. Yang. Dip: Dual incongruity perceiving network for sarcasm detection. In *CVPR*, 2023. 2

[68] C. Wen, X. Zhang, X. Yao, and J. Yang. Ordinal label distribution learning. In *ICCV*, 2023. 2

[69] H. Wu, Z. Yao, J. Wang, and M. Long. Motionrnn: A flexible model for video prediction with spacetime-varying motions. In *CVPR*, 2021. 2

[70] J. Xing, W. Hu, Y. Zhang, and T.-T. Wong. Flow-aware synthesis: A generic motion model for video frame interpolation. *CVMJ*, 2021. 1

[71] B. Yang, C. Liu, B. Li, J. Jiao, and Q. Ye. Prototype mixture models for few-shot semantic segmentation. In *ECCV*, 2020. 4

[72] R. Yang, P. Srivastava, and S. Mandt. Diffusion probabilistic modeling for video generation. *arXiv*, 2022. 2, 6, 7

[73] S. Yang, L. Zhang, Y. Liu, Z. Jiang, and Y. He. Video diffusion models with local-global context guidance. In *IJCAI*, 2023. 2

[74] X. Ye and G.-A. Bilodeau. Vptr: Efficient transformers for video prediction. *ICPR*, 2022. 2

[75] X. Ye and G.-A. Bilodeau. A unified model for continuous conditional video prediction. In *CVPR*, 2023. 2

[76] L. Yu, Y. Cheng, K. Sohn, J. Lezama, H. Zhang, H. Chang, A. G. Hauptmann, M.-H. Yang, Y. Hao, I. Essa, et al. Magvit: Masked generative video transformer. *arXiv*, 2022. 5

[77] S. Yu, K. Sohn, S. Kim, and J. Shin. Video probabilistic diffusion models in projected latent space. In *CVPR*, 2023. 2

[78] W. Yu, Y. Lu, S. Easterbrook, and S. Fidler. Efficient and information-preserving future frame prediction and beyond. In *ICLR*, 2020. 2

[79] Y. Zhai, G. Jia, Y.-K. Lai, J. Zhang, J. Yang, and D. Tao. Looking into gait for perceiving emotions via bilateral posture and movement graph convolutional networks. *TAC*, 2024. 1

[80] P. Zhang, D. Wang, and H. Lu. Multi-modal visual tracking: Review and experimental comparison. *CVMJ*, 2024. 1

[81] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[82] Z. Zhang, S. Chen, Z. Wang, and J. Yang. Planeseg: Building a plug-in for boosting planar region segmentation. *TNNLS*, 2024. 1

[83] Z. Zhang, J. Hu, W. Cheng, D. Paudel, and J. Yang. Extdm: Distribution extrapolation diffusion model for video prediction. In *CVPR*, 2024. 1

[84] Z. Zhang, S. Liu, and J. Yang. Multiple planar object tracking. In *ICCV*, 2023. 1

[85] Z. Zhang, L. Wang, and J. Yang. Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network. In *CVPR*, 2023. 1

[86] Z. Zhang and J. Yang. Temporal sentiment localization: Listen and look in untrimmed videos. In *ACM MM*, 2022. 1

[87] Z. Zhang, P. Zhao, E. Park, and J. Yang. Mart: Masked affective representation learning via masked temporal distribution distillation. In *CVPR*, 2024. 2

[88] P. Zhao, P. Xu, P. Qin, D.-P. Fan, Z. Zhang, G. Jia, B. Zhou, and J. Yang. Lake-red: Camouflaged images generation by latent background knowledge retrieval-augmented diffusion. In *CVPR*, 2024. 2

[89] S. Zhao, G. Jia, J. Yang, G. Ding, and K. Keutzer. Emotion recognition from multiple modalities: Fundamentals and methodologies. *SPM*, 2021. 2

[90] S. Zhao, X. Yao, J. Yang, G. Jia, G. Ding, T.-S. Chua, B. W. Schuller, and K. Keutzer. Affective image content analysis: Two decades review and new perspectives. *TPAMI*, 2022. 2

[91] S. Zhou, D. Chen, J. Pan, J. Shi, and J. Yang. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *CVPR*, 2024. 2

[92] S. Zhou, M. Jiang, S. Cai, and Y. Lei. Dc-gnet: Deep mesh relation capturing graph convolution network for 3d human shape reconstruction. In *ACM MM*, 2021.

[93] S. Zhou, M. Jiang, Q. Wang, and Y. Lei. Towards locality similarity preserving to 3d human pose estimation. In *ACCV*, 2020. 2

[94] X. Zhu, J. Song, L. Gao, F. Zheng, and H. T. Shen. Unified multivariate gaussian mixture for efficient neural image compression. In *CVPR*, 2022. 4