

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(XXXX)XX-0001-14

论文引用格式: Chen Yiren, Zhu Changwen, Zhao Jiawei, Liu Deyang, Zha Hongbin, Shan Caifeng. Spatio-temporal alignment diffusion prediction model for embryo development[J/OL]. Journal of Image and Graphics, XXXX: 1-14. DOI: 10.11834/jig.260255. (陈义仁, 朱昌文, 赵家培, 刘德阳, 查红彬, 单彩峰. 面向胚胎发育的时空对齐扩散预测模型[J/OL]. 中国图象图形学报, XXXX: 1-14. DOI: 10.11834/jig.260255.) [DOI: 10.11834/jig.260255]

面向胚胎发育的时空对齐扩散预测模型

陈义仁¹, 朱昌文¹, 赵家培¹, 刘德阳^{1*}, 查红彬^{1,2}, 单彩峰³

1. 安庆师范大学人工智能与计算机学院, 安庆 246133; 2. 北京大学机器感知与智能教育部重点实验室, 北京 100871; 3. 南京大学智能科学与技术学院, 苏州 215163

摘要: 目的 在体外受精(in vitro fertilization, IVF)发育预测任务中, 胚胎发育过程具有高度非线性的形变和显著的个体差异, 这使得长跨度预测中易出现结构模糊, 从而降低生成序列的视觉保真度。因此, 本研究提出时空对齐扩散模型(spatio-temporal alignment diffusion model, STA-Diff), 以增强胚胎从卵裂期到囊胚期演化过程的预测稳定性。方法 STA-Diff 首先通过运动表征重建器(motion representation reconstructor, MRR)建立图像域与运动空间之间的非线性映射, 将预测过程由像素空间转移至运动表征空间, 实现外观信息与发育动态特征的解耦, 降低冗余纹理对长期演化建模的影响。在此基础上, 针对运动空间中胚胎发育过程存在的局部结构变化、长时序依赖以及个体发育速率差异问题, STA-Diff 分别引入标准化帧间残差先验(standardized inter-frame residual prior, SIRP)、动态时空解耦门控混合器(dynamic spatio-temporal decoupling mixer, D-STDM)和时间加权条件调制(temporal-weighted conditional modulation, TWCM), 分别强化关键结构变化区域的运动约束、建模跨阶段发育过程中的长程依赖关系, 将运动表征与生成特征的动态对齐, 从而提升长序列预测过程中的结构稳定性和视觉一致性。结果 在公开 IVF 延时成像数据集上的实验表明, STA-Diff 在多项指标上均取得最优结果, PSNR(peak signal-to-noise ratio)达 23.85dB, SSIM(structural similarity index, SSIM)达 0.6542, LPIPS(learned perceptual image patch similarity)较主流方法降低约 55% - 75%。关键结构区域的 Dice 系数(Dice coefficient)达到 0.936。结论 STA-Diff 改善了胚胎发育的长序列图像生成质量, 使模型可在早期卵裂期准确预测囊胚期形态, 为 IVF 临床中的前瞻性胚胎质量评估提供了一种可靠的方法。代码链接 <https://www.scidb.cn/s/e6ry6z>。

关键词: 胚胎发育预测; 时空对齐; 扩散模型; 运动表征; 特征解耦

Spatio-temporal alignment diffusion prediction model for embryo development

Chen Yiren¹, Zhu Changwen¹, Zhao Jiawei¹, Liu Deyang^{1*}, Zha Hongbin^{1,2}, Shan Caifeng³

1. School of Artificial Intelligence and Computer Science, Anqing Normal University, Anqing 246133, China; 2. Key Laboratory of machine perception and intelligence, Ministry of Education, Peking University, Beijing 100871, China; 3. School of Intelligence Science and Technology, Nanjing University, Suzhou 215163, China

Abstract: Objective In vitro fertilization (IVF) is an assisted reproductive technology where mature oocytes are fertilized

收稿日期: 2026-05-07; 修回日期: 2026-08-26

* 通信作者: 刘德阳 deyang.liu@hotmail.com

基金项目: 国家自然科学基金(62171002); 安徽省自然科学基金(2508085Y039); 安徽省高校杰出青年科研项目(2024AH020009); 安徽省高端人才引育行动项目

Supported by: National Natural Science Foundation of China (62171002); Natural Science Foundation of Anhui Province (2508085Y039); Outstanding Youth Research Project of Anhui Provincial Universities (2024AH020009); High-end Talent Introduction and Cultivation Project of Anhui Province

outside the body. Then, they are cultured under controlled conditions for approximately one week to monitor embryonic development. Embryos with the highest developmental potential are transferred back into the uterus to maximize the chances of a successful pregnancy. With the widespread clinical application of time-lapse imaging (TLI) technology in IVF, automatic analysis of embryo image sequences has gradually become an important research direction for assisting embryo selection. Existing studies predominantly employ deep learning techniques and generally follow a retrospective modeling paradigm, focusing on tasks such as embryo quality assessment and developmental stage classification. These approaches are mainly oriented toward discriminative analysis or scoring of observed developmental trajectories, rather than toward explicit modeling of future morphological evolution. To date, limited attention has been paid to the prospective generative prediction of long-term, nonlinear developmental progression from the cleavage stage to the blastocyst stage. Although existing spatiotemporal prediction models can generate future image sequences based on historical observations, they often lack explicit constraints for modeling cross-stage structural transitions, particularly in biological developmental sequences characterized by pronounced stage shifts and structural reorganization. Directly applying these models to long-term prediction tasks, such as embryonic development forecasting, poses significant challenges. These difficulties stem primarily from two factors: the strong nonlinearity of embryonic morphological deformation, and the marked inter-individual variability in developmental rates. Both factors are highly prone to inducing structural ambiguity in long-range prediction scenarios. With the prolongation of the temporal horizon, the accumulation of recursive errors intensifies this ambiguity, thereby resulting in a marked decline in the visual fidelity of the generated sequences. To address these limitations, we propose the spatio-temporal alignment diffusion model (STA-Diff), a comprehensive framework designed to enhance the structural stability and predictive consistency of embryonic evolution from the cleavage stage to the blastocyst stage. **Methods** STA-Diff facilitates high-fidelity embryonic forecasting by leveraging the motion representation reconstructor (MRR) to characterize the nonlinear mapping between the latent manifold of the image domain and the underlying motion field. By projecting the prediction targets in the high-dimensional pixel space onto a canonical and stable motion representation space, the framework effectively decouples static appearance information from dynamic motion features. This transformation effectively reduces appearance redundancy and alleviates the structural degradation inherent in long-term evolution modeling. In this latent motion representation space, STA-Diff carries out sequence synthesis by means of a conditional diffusion probabilistic process. Since the evolution of local structures during embryonic development exhibits significant temporal non-uniformity—that is, marked variations in morphological changes across different stages—this heterogeneous distribution may lead to an uneven allocation of denoising capability if not explicitly modeled. To this end, we introduce the standardized inter-frame residual prior (SIRP), which provides region-aware structured guidance to balance the denoising process between rapid transitions and stable regions in long-term scenarios. Furthermore, maintaining spatiotemporal consistency is crucial for characterizing continuous embryonic dynamics when capturing long-range dependencies across multiple stages. In view of the fact that conventional spatiotemporal fusion architectures are frequently inadequate for coping with such intricate transformations, we introduce the dynamic spatio-temporal decoupling mixer (D-STDm). This module explicitly decouples spatial modeling from temporal propagation and integrates them via a content-adaptive gating mechanism, thereby significantly enhancing the modeling of long-range temporal dependencies. Moreover, in view of the fact that the contribution of conditional information varies over distinct temporal segments, we propose the temporal-weighted conditional modulation (TWCM) module. The TWCM module does not apply uniform weighting to all conditional information, as this might suppress critical developmental signals. Instead, it adaptively reweights the conditional information along the temporal dimension. These aggregated representations are then injected into the denoising network via affine modulation based on feature-wise linear modulation (FiLM), thereby enabling selective and time-aware guidance throughout the diffusion process. **Results** Experimental evaluations on a public IVF time-lapse imaging dataset demonstrate that STA-Diff consistently outperforms 13 state-of-the-art spatiotemporal prediction methods across multiple benchmarks. On the blastocyst-stage prediction task, our method attains a peak signal-to-noise ratio (PSNR) of 23.85 dB and an structural similarity index (SSIM) of 0.6542. Moreover, it reduces the learned perceptual image patch similarity (LPIPS) by approximately 55%–75% compared to the state-of-the-art spatiotemporal prediction approaches. In addition, the Dice coefficient for key structural regions reaches 0.936, indicating that the proposed method effectively alleviates the

morphological ambiguity commonly encountered in long-term prediction. **Conclusion** The proposed model integrates the inherent stability of motion representation modeling with the powerful generative capabilities of diffusion models. This combination leads to a marked enhancement in the quality of generated predictions for long-term embryonic development. Compared to conventional methods, our framework maintains robust temporal consistency and structural integrity over extended time horizons. It enables accurate prediction and generation of the complete blastocyst-stage morphology and spatial details, based directly on observations from the early cleavage stage. These results not only enhance the visual fidelity and physiological plausibility of long-term embryo development prediction, but also provide stable, reliable, and quantifiable technical support for clinical IVF applications, such as early embryo tracking, dynamic assessment, and prospective quality screening. The codes are available at: <https://www.scidb.cn/s/e6ry6z>.

Key words: embryo development prediction; spatio-temporal alignment; diffusion models; motion representation; feature disentanglement

论文引用格式: Chen Yiren, Zhu Changwen, Zhao Jiawei, Liu Deyang, Zha Hongbin, Shan Caifeng. Spatio-temporal alignment diffusion prediction model for embryo development [J/OL]. Journal of Image and Graphics, XXXX: 1-13. DOI: 10.11834/jig.260255. (陈义仁, 朱昌文, 赵家培, 刘德阳, 查红彬, 单彩峰. 面向胚胎发育的时空对齐扩散预测模型[J/OL]. 中国图象图形学报, XXXX: 1-13. DOI: 10.11834/jig.260255.)

论文引用格式: Chen Yiren, Zhu Changwen, Zhao Jiawei, Liu Deyang, Zha Hongbin, Shan Caifeng. Spatio-temporal alignment diffusion prediction model for embryo development [J/OL]. Journal of Image and Graphics. DOI: 10.11834/jig.260255. (陈义仁, 朱昌文, 赵家培, 刘德阳, 查红彬, 单彩峰. 面向胚胎发育的时空对齐扩散预测模型[J/OL]. 中国图象图形学报. DOI: 10.11834/jig.260255)

0 引言

体外受精 (in vitro fertilization, IVF) 是一项广泛应用的辅助生殖技术。该治疗过程首先从女性体内采集一定数量的成熟卵子, 并在体外完成受精, 形成受精卵或胚胎并在培养箱中继续生长一周。医生会筛选出质量最好的胚胎, 将其移植回女性胚胎内期实现妊娠 (Inhorn 等, 2015; Araya 等, 2024; Barriere 等, 2018)。筛选评估每个胚胎质量并对所有可用胚胎进行排序的过程称为胚胎评估。

随着延时成像技术在体外受精 (IVF) 临床中的广泛应用, 基于胚胎图像序列的自动化分析逐渐成为辅助胚胎筛选的重要研究方向。现有的研究多采

用深度学习技术 (Krizhevsky 等, 2012; Dosovitskiy 等, 2020; Zhang 等, 2025), 对胚胎研究主要集中于胚胎质量评估、发育阶段识别以及妊娠或着床结局预测, 从单帧图像或短时序片段中提取判别性特征。这类方法在减少人工评估主观性、提升分级一致性方面取得了一定进展 (Guo 等, 2023; Wu 等, 2021; Dirvanauskas 等, 2019; Liu 等, 2025), 如 Kragh 等人 (2021) 使用自监督对齐对胚胎质量进行评估。

然而, 上述研究大多遵循回顾性建模范式, 其核心假设是胚胎的关键发育阶段已被完整观测, 模型目标在于对既有发育轨迹进行判别与评分, 而非对未来发育阶段的形态演化进行显式建模。

在计算机视觉领域, 时空预测 (Lotter 等, 2016; Gao 等, 2022; Dai 等, 2019; Wang 等, 2022) 旨在根据历史观测帧生成未来图像序列。早期时空预测研究以传统非扩散方法为主, 可分为循环型与非循环型两大类。其中, 循环型方法的核心特征是显式引入递归神经网络结构 (Shi 等, 2015; Wang 等, 2019; Tang 等, 2023; Chang 等, 2021; Wang 等, 2019; Le 等, 2020)。非循环型方法 (Gao 等, 2022; Tan 等, 2023; Zhong 等, 2023; Nie 等, 2024) 采用编码器、解码器架构, 直接对输入的历史帧序列进行全局时空特征提取, 并一步到位地输出未来所有帧, 避免了循环结构在长序列预测中易出现的误差累积问题。

近年来, 扩散式生成模型 (Ho 等, 2020) 被引入时空预测, 通过建模条件概率分布更好地刻画未来演化的不确定性 (Höppe 等, 2022; Zhang 等, 2024)。然而, 现有扩散式视频预测方法多数直接在外观或高维特征空间对未来帧进行分布建模, 并采用统一的时间条件注入方式。这类设计在自然视频场景中表现有效, 直接应用于胚胎发育等长跨度预测任务

时,模型仍面临结构一致性保持和条件时序对齐等方面的挑战。

基于上述背景,本研究构建时空对齐扩散模型(spatio-temporal alignment diffusion model, STA-Diff),将运动解耦与扩散生成相结合,通过结构引导、时序建模与条件对齐三类机制,有效缓解长跨度预测中的结构模糊与误差累积,实现早期卵裂期对囊胚期形态的前瞻性预测。

1 本文方法

1.1 总体框架

本研究面向胚胎发育的长跨度阶段预测任务,给定同一胚胎在 T_i 个阶段的观测序列 $I_{1:T_i} = \{I_1, \dots, I_{T_i}\}$,生成随后 T_o 个阶段囊胚阶段的预测序列

$\hat{I}_{(T_i+1:T_i+T_o)} = \{\hat{I}_{T_i+1}, \dots, \hat{I}_{T_i+T_o}\}$ 。在像素空间进行多步预测时易产生细节模糊与误差累积。为此,本研究提出 STA-Diff(总体框架如图 1 所示),采用两阶段建模策略以提升长跨度预测的稳定性。STA-Diff 的第一阶段利用运动表征重建器学习图像帧与运动场之间的映射关系,构建运动表征空间。在此基础上,第二阶段通过条件运动表征扩散预测模型(conditional motor representation diffusion model, CMRDM),在运动空间中推未来囊胚阶段的运动表征。CMRDM 在扩散建模过程中结合多种针对性设计以应对胚胎发育的复杂变化。最终,预测得到的表征经运动表征重建器(motion representation reconstructor, MRR)解码器映射回图像空间,生成胚胎的预测阶段图像。

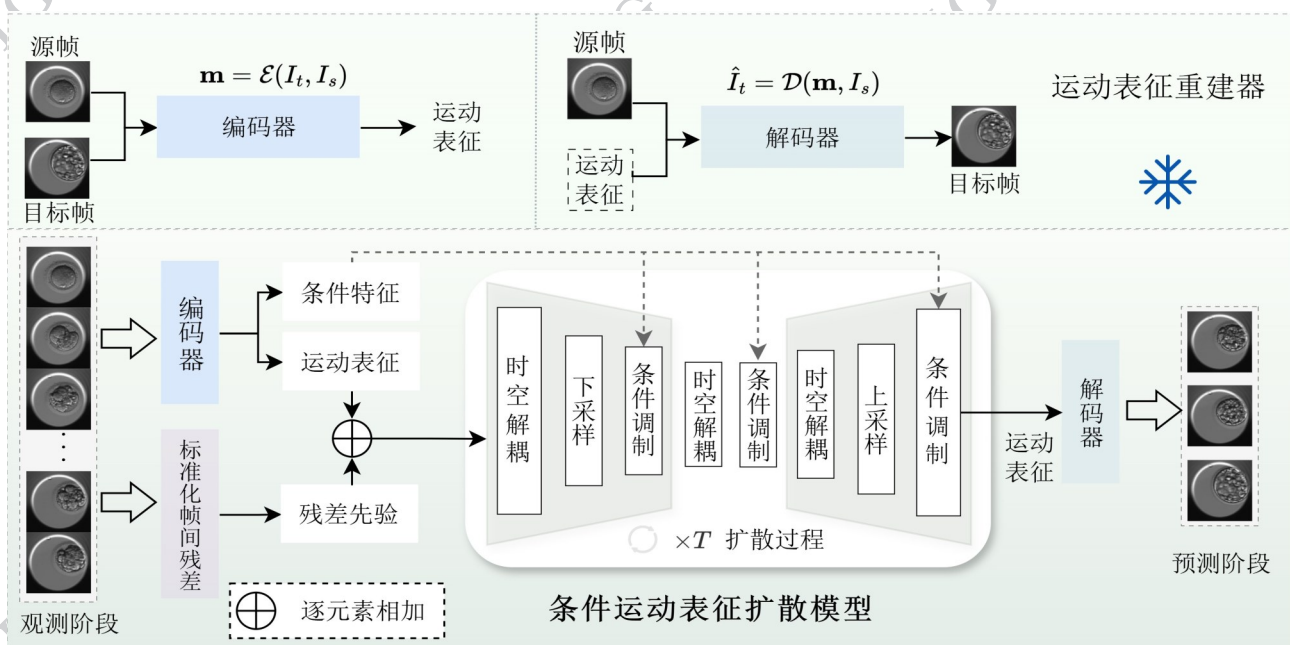


图 1 STA-Diff 整体框架图

Fig. 1 Overall framework diagram of STA-Diff

1.2 运动表征重建器

MRR 的作用在于学习图像帧与运动表征之间的近似双向映射,将预测问题从易受外观变化影响的像素空间转移至更适合长期建模的运动表征空间。

模块由编码器 $\mathcal{E}$ 与解码器 $\mathcal{D}$ 组成:编码器从成对帧 (I_t, I_s) 中提取运动表征 m ,解码器在给定源帧 I_t 与运动表征 m 的条件下重建目标帧 $\hat{I}_s$ 。

在 MRR 训练阶段,本研究从同一胚胎序列中随机采样任意两个阶段,构造源阶段图像 I_t 与目标阶段图像 I_s 。编码器 $\mathcal{E}$ 通过成对输入估计两帧之间的运动相关性,并输出由光流与置信图组成的运动表征,其形式定义为:

$$m = \mathcal{E}(I_t, I_s), \quad m = [U; C] \quad (1)$$

式中 $U \in \mathbb{R}^{2 \times h \times w}$ 表示从 I_t 到 I_s 的稠密形变场; $C \in [0, 1]^{1 \times h \times w}$ 为置信图。

在重建环节中, 解码器 $\mathcal{D}$ 接收参考帧 I_s 与运动表征 m 为输入, 通过特征对齐与精修生成目标阶段的重建图像 $\hat{I}_t$, 具体表达式如下:

$$\hat{I}_t = \mathcal{D}(m, I_s) \quad (2)$$

1.3 条件运动表征扩散模型

在 MRR 构建的运动表征空间基础上, 本研究将长跨度阶段预测任务转化为对未来运动表征序列的条件外推建模。CMRDM 采用标准扩散建模框架, 通过前向加噪与条件反向去噪过程学习未来运动表征的条件概率分布。然而直接应用条件扩散模型面临三方面挑战: 局部结构变化不均匀, 缺乏先验约束易导致生成区域与实际发育规律不一致; 从卵裂期到囊胚期的阶段转换复杂, 需同时捕获局部形态变化与跨阶段时序依赖; 不同胚胎发育速率的差异导致条件信息与目标状态之间会出现时间语义偏移。

针对上述问题, 本研究在 CMRDM 中分别引入三种对应机制: 标准化帧间残差先验 (standardized inter-frame residual prior, SIRP) 应对局部结构变化不均匀问题, 提供区域变化引导; 动态时空解耦门控混合器 (dynamic spatio-temporal decoupling mixer, D-STDM) 增强了对跨阶段时序依赖关系的建模能力; 时间加权条件调制 (temporal-weighted conditional modulation, TWCM) 应对时间语义偏移问题, 实现条件信息的自适应时间对齐。三者协同作用, 提升长序列预测的结构稳定性和视觉一致性。

1.3.1 标准化帧间残差先验

胚胎发育过程中局部结构演化存在显著的时序不均匀性, 不同阶段间的形态变化幅度差异较大。为此, 本研究引入 SIRP, 为扩散预测提供区域级结构变化引导。SIRP 首先对条件阶段序列中相邻帧进行灰度转换, 计算帧间灰度残差 R_t 以刻画局部结构变化强度。为消除不同胚胎样本间整体亮度与成像条件差异的影响, 对残差序列进行标准化处理, 具体表达式如下:

$$\tilde{R}_t = \frac{R_t}{\sigma(R) + k}, \quad \sigma(R) = \sqrt{\frac{1}{T_t - 1} \sum_{i=1}^{T_t-1} (R_i - \bar{R})^2} \quad (3)$$

式中 $\bar{R}$ 为残差序列的均值, $\sigma(R)$ 为残差序列标准差, $k = 10^{-5}$ 为数值稳定性常数。标准化后的残差 $\tilde{R}_t$ 被下采样至与运动表征一致的空间分辨率, 并作为

附加结构先验与条件特征一同注入扩散去噪网络。

1.3.2 动态时空解耦门控混合器

针对长跨度运动表征预测任务, 模型需兼顾局部空间结构的时空一致性与多发育阶段间的时序依赖关系。因此本研究引入动态时空解耦门控混合器 D-STDM, 通过显式分离空间建模与时间传播过程, 并利用内容自适应的门控机制对两类特征进行融合, 从而增强对长程时序依赖的建模能力。

在 CMRDM 的去噪过程中, 扩散网络每一层接收的运动表征特征为 $F \in \mathbb{R}^{b \times c \times t \times h \times w}$, b, c, t, h, w 分别表示批大小、通道数、预测阶段数量及空间分辨率。如图 2 所示, D-STDM 首先将特征 F 在时间维度上展开为二维特征 $F_{2d} \in \mathbb{R}^{(b \cdot c) \times t \times h \times w}$, 并经由二维卷积、归一化与非线性激活函数组成的空间混合网络进行处理, 得到空间混合特征。最后将混合特征恢复为原始时序结构, 得到空间分支输出 $F_s \in \mathbb{R}^{b \times c \times t \times h \times w}$ 。

时间分支用于建模跨阶段的演化依赖关系。以空间混合后的特征 F_s 为输入, 沿时间维度对特征进行递归传播。然后对第 t 个时间步的特征切片 $F_{s,t}$ 采用卷积门控循环单元 (convolutional gated recurrent unit, ConvGRU) 进行递归更新, 随后将所有时间步的隐藏状态沿时间维拼接, 得到时间分支的输出 F_t 。

为实现空间与时间特征的自适应融合, D-STDM 进一步引入门控机制。空间与时间特征门控权重由空间分支特征 F_s 通过三维卷积网络预测得到, 空间与时间分支特征通过门控方式进行融合, 再以残差连接形式与原始特征相加得到加权的特征 $\tilde{F}_0$ 。

1.3.3 时间加权条件调制

在胚胎发育预测任务中, 不同时间片段的条件信息对当前去噪过程的贡献程度并不一致, 等权处理会削弱关键阶段的作用。为此, 本研究引入 TWCM, 在时间维度上对条件信息自适应加权, 并以仿射调制 (feature-wise linear modulation, FiLM) 注入去噪网络, 实现对扩散过程的选择性引导。

在 TWCM 中, 整体结构如图 3 所示, 本研究以当前层的运动表征特征 $\tilde{F}_0$ 作为调制对象, 并以 MRR 编码器中条件特征 F_t 提供外观与阶段语义约束。为评估不同时间片段条件信息的相对重要性, 首先对特征空间维进行平均池化, 得到帧级向量 f_i 与 f_i^c 。

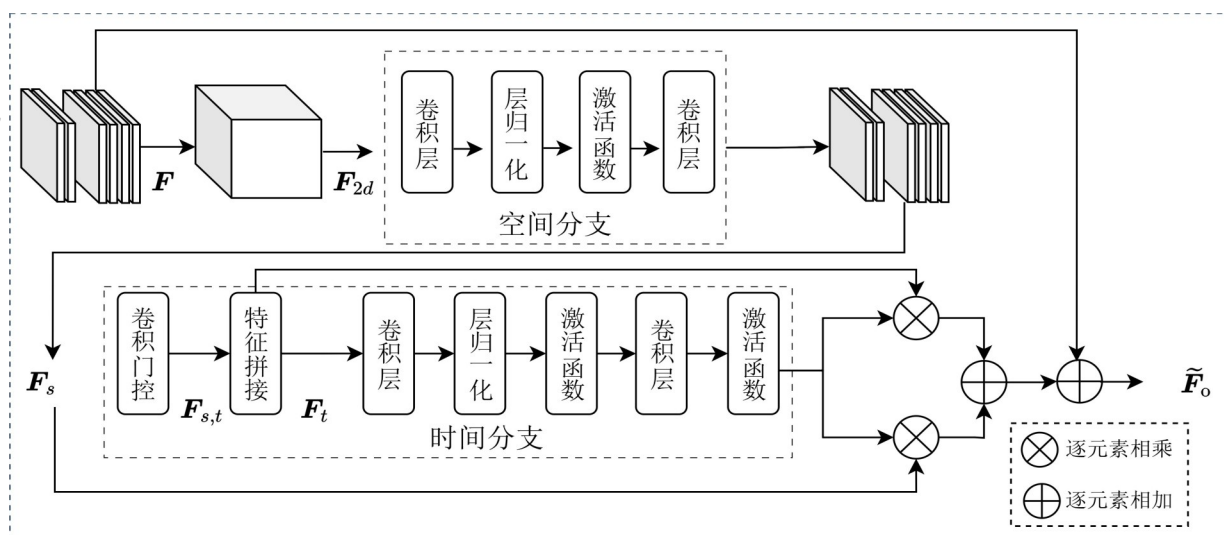


图2 D-STDM网络结构图

Fig. 2 Architecture diagram of D-STDM

随后,帧级特征 f_i 与 f_i^c 分别投影到共享嵌入空间,并基于逐帧余弦相似度评估二者在时间维上的一致性。相似度在时间维上归一化后得到时间权重 w_t ,用于刻画每个时间片段条件信息对当前生成过程的相对贡献。基于权重 w_t ,TWCM在时间维上对条件

帧级表示进行加权汇聚,得到全局条件向量 f_a^c 。最后,聚合后的条件向量 f_a^c 被映射为通道级FiLM仿射参数 γ 与 β ,并以其共享调制的方式注入到 $\tilde{F}_o$ 中,实现对扩散去噪过程的条件调制。

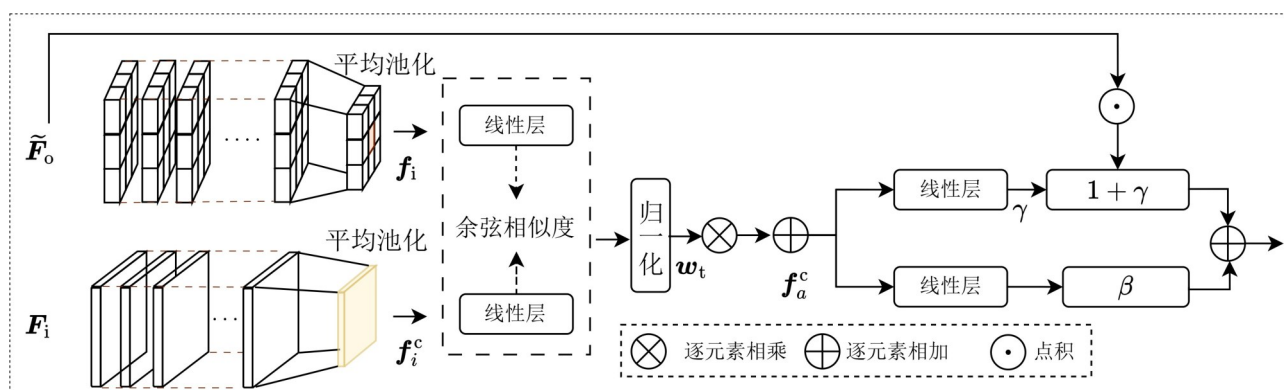


图3 TWCM网络结构图

Fig. 3 Architecture diagram of TWCM

2 实验

2.1 数据集

本研究的实验验证基于法国南特大学(Gomez等,2022)公开的IVF胚胎延时成像发育数据集。本研究从中筛选出完整覆盖15个标准发育阶段的117个独立胚胎序列作为原始样本。对每个胚胎的15个发育阶段,在对应时间窗口内随机采样一帧图像,在保持阶段顺序的前提下构建多组时序序列,每个

原始序列可生成20组形态细节有差异但阶段语义一致的样本。

117个胚胎被划分为训练集97个、验证集10个、测试集10个,增强后训练集包含1940组样本,验证集和测试集各200组。

2.2 评价指标

本研究从图像层面和区域层面两个尺度评价预测质量,所有指标在12个输入阶段后生成的3个囊胚期预测帧上逐帧计算,并在时间维和样本维取均值。图像层面采用均方误差(mean square error,

MSE)、峰值信噪比(peak signal-to-noise ratio, PSNR)(Wang等, 2009)、结构相似性指数(structural similarity index, SSIM)(Wang等, 2004)和感知相似度指标(learned perceptual image patch similarity, LPIPS)(Zhang等, 2018), 分别衡量像素分布、结构信息和感知一致性。区域层面将人工标注的关键形态区域转换为二值掩码, 利用预训练微调的U-Net(Ronneberger等, 2015)对生成图像进行区域提取, 采用交并比(intersection over union, IoU)与Dice系数(Dice coefficient, Dice)评估空间重叠程度, 并计算预测区域与真实区域的面积比例以衡量结构尺度预测准确性。

2.3 实验结果

2.3.1 整体图像生成质量分析

图4展示了不同方法在囊胚阶段的预测结果。多数基于像素空间建模的方法在预测后期存在不同程度的模糊或结构退化, 而STA-Diff能够保持较为清晰的结构轮廓和稳定的形态演化趋势。

表1汇总了定量比较结果。现有方法在PSNR上多分布于约22dB, 较优者不超过23.1dB; SSIM主要集中在0.58 - 0.63区间。STA-Diff的PSNR达

23.85dB, 较现有方法提高约0.7 - 1.8dB; SSIM达0.6542, 提升约3% - 7%。在LPIPS上, 多数对比方法高于0.13, 而STA-Diff降低至0.0571, 实现了约55% - 75%的误差降低, 显示出在感知质量与结构细节保持方面的明显优势。

2.3.2 关键运动区域定量分析

在胚胎发育过程中, 形态演化的核心动态主要集中在局部区域。相较于全图评价, 关键运动区域的定量分析能够更直接反映模型对胚胎发育关键结构演化过程的建模能力。因此本研究在囊胚期阶段对不同方法的预测结果进行了区域层面的定量评估, 相关可视化结果如图5所示。多数基于像素空间建模的时序预测方法在胚胎关键运动区域仅能保持基本的纹理结构, 但在精细结构、边缘轮廓与边界清晰度上的预测效果仍较为模糊。

表2汇总了定量结果。现有方法在区域层面的性能普遍低于全图评价, SSIM集中于0.28 - 0.37, LPIPS多高于0.09。STA-Diff的区域SSIM达0.3985, 提升约8% - 40%; LPIPS降0.0534, 降低约40% - 70%, 表明其预测结果在细节建模方面和对比方法相比更优。

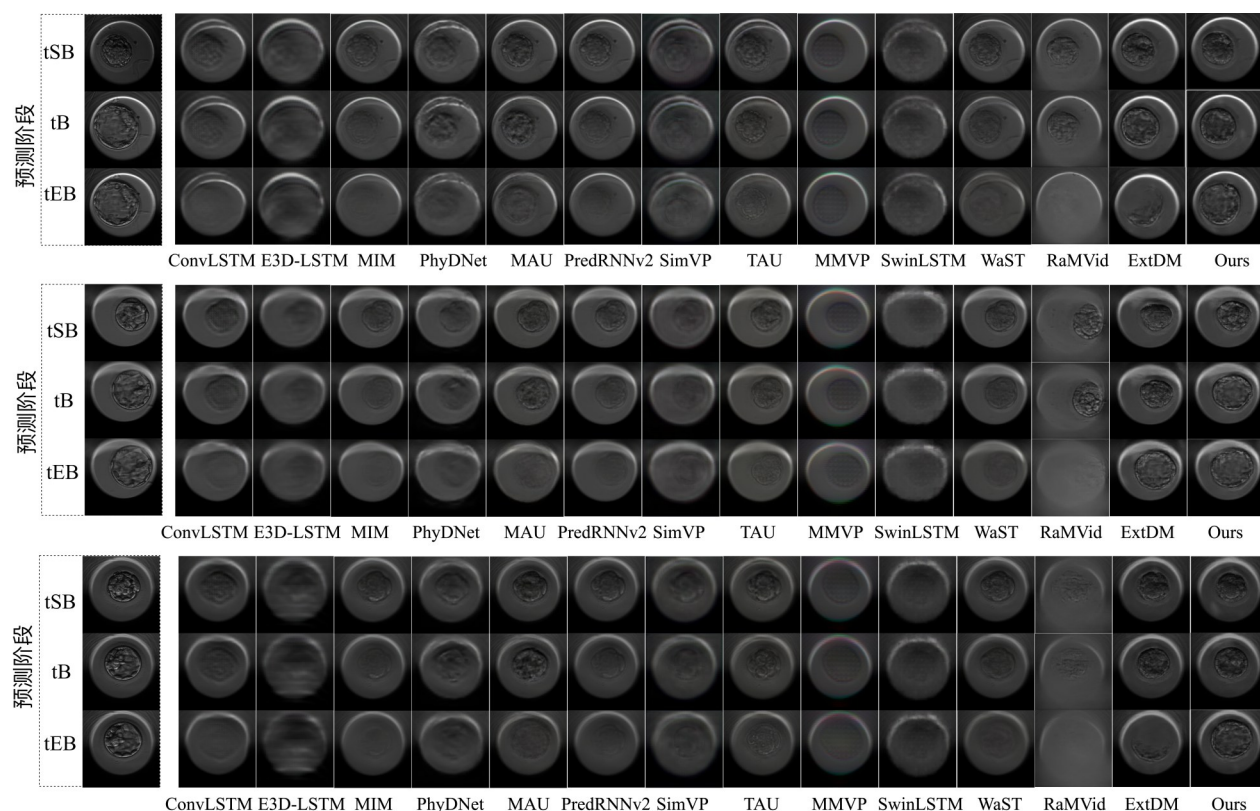


图4 STA-Diff与不同预测方法囊胚期完整图像对比

Fig. 4 Comparison of blastocyst images between STA-Diff and other prediction methods

表 1 不同方法预测囊胚期图像的定量指标对比
Table 1 Quantitative comparison of images at blastocyst stage predicted by different methods

Model	MSE ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
ConvLSTM(Shi 等, 2015)	0.00555 5	22.55	0.6137	0.1707
E3D-LSTM(Wang 等, 2019)	0.00466 8	23.31	0.6071	0.2491
MIM(Wang 等, 2019)	0.00536 3	22.71	0.6225	0.1707
PhyDNet(Le 等, 2020)	0.00575 5	22.4	0.5417	0.1973
MAU(Chang 等, 2021)	0.00674 1	21.71	0.5828	0.1276
PredRNNv2(Wang 等, 2021)	0.00490 5	23.09	0.6354	0.1457
SimVP(Gao 等, 2022)	0.00536 8	22.7	0.5658	0.2307
TAU(Tan 等, 2023)	0.00575 5	22.4	0.6109	0.1409
MMVP(Zhong 等, 2023)	0.00682 5	21.66	0.5323	0.2665
SwinLSTM(Tang 等, 2023)	0.00581 7	22.35	0.5922	0.2791
WaST(Nie 等, 2024)	0.00500 1	23.01	0.6279	0.1353
RaMViD(Höppe 等, 2022)	0.02471 3	16.07	0.2824	0.3935
ExtDM(Zhang 等, 2024)	0.00713 6	21.47	0.6022	0.0978
STA-Diff	0.00412	23.85	0.6542	0.0571

注:黑色字体表示最优结果。↓表示值越小越好;↑表示值越大越好。

2.3.3 复杂形态边界的预测可靠性研究

本研究基于预训练并少样本微调的U-Net对囊胚期图像进行关键区域提取,计算区域重叠与面积指标,结果汇总于表3。在区域重叠指标上,多数对比方法在Dice与IoU上相较于全图结果出现不同程度下降,STA-Diff的Dice与IoU分别达0.936和0.8798,处于现有方法性能分布的最高值,表明其在关键结构区域的空间一致性保持方面具有更高可靠性。

由于胚胎发育过程中囊胚区域面积会随时间发生动态变化,本研究进一步分析预测区域与真实区域之间的面积比例,以评估模型对胚胎结构尺度变化及发育速度的预测能力。该指标为预测区域面积与真实区域面积的比值,其理想值为1.0。当比例大于1.0时,表示模型预测区域扩张速度快于真实发育过程;当比例小于1.0时,则表示模型预测区域扩张滞后。STA-Diff的区域面积比例为1.0701,显示其在关键结构尺度预测过程中能够较好地维持区域扩张与收缩趋势的稳定性。图6展示了不同方法在预测过程中的区域预测稳定性变化,红色为真实囊胚区域,绿色为U-Net识别区域。预测初期仅少数方法失败;随着预测时间的逐步推进,失败方法数量明显增加,最远期预测阶段约三分之二的对比方法Dice与IoU指标接近于零,反映出这些方法在长跨度预测中难以维持稳定的结构变化。

2.3.4 消融实验

为评估各模块贡献,本研究通过逐步移除SIRP、D-STDM及TWCM进行消融实验,定量结果汇总于表4。

移除SIRP后,模型在囊胚期预测中出现局部结构异常(如图7红色箭头所示的白色伪影),SSIM下降3.41%,LPIPS上升21.72%,表明缺乏结构先验时扩散去噪难以聚焦关键发育区域。

在进一步同时移除SIRP与D-STDM模块的情况下,模型预测结果呈现出发育进程滞后现象。预测生成的胚胎形态在囊胚阶段仍保持接近早期卵裂阶段的外观特征,未能形成囊胚期特有的形态重构与扩张变化,相较于仅移除SIRP,PSNR指标下降0.62,反映出模型在缺乏长程时空依赖建模时难以维持跨阶段的连续形态演化。模型参数量计算量虽降低至51.5 M,1081G浮点运算次数(floating point operations, FLOPs),但也说明了D-STDM模块在增加一定计算成本的同时,为模型提供了有效的时空动态建模能力。

分别保留单一模块时,仅使用SIRP时SSIM降至0.5938;仅保留D-STDM获得PSNR 22.57dB、SSIM 0.6027,说明D-STDM对长期时序建模作用更明显,SIRP主要提供结构变化约束。

进一步移除TWCM后,参数量虽然降低0.43M,但生成图像亮度与对比度下降,结构边界模糊,LPIPS上升0.0151,整体视觉质量受到明显影响。

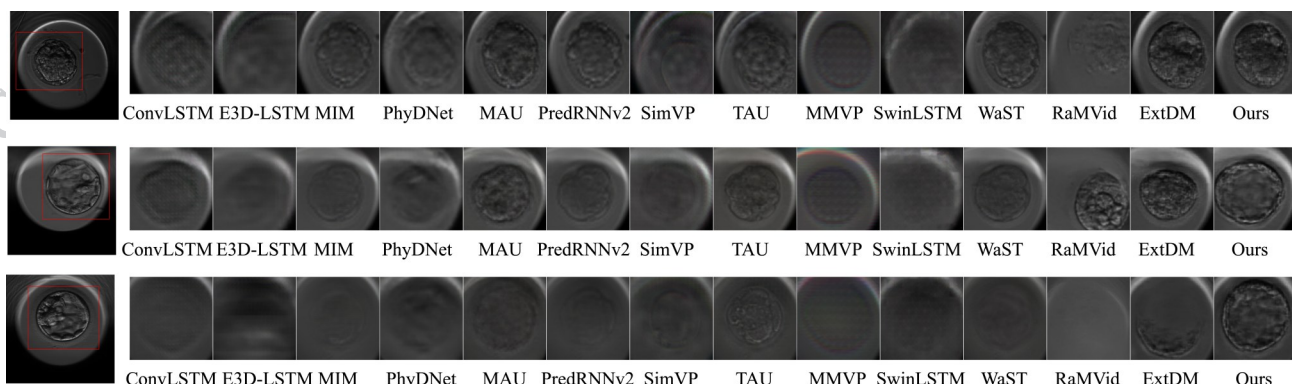


图5 STA-Diff不同视频预测方法囊胚关键运动区域对比

Fig. 5 Comparison of key motion regions in blastocysts between STA-Diff and other video prediction methods

综上,各模块逐步移除对时序一致性、空间结构完整性产生叠加影响,验证了STA-Diff中各核心模块的必要性与协同作用。

3 讨论

3.1 运动与外观解耦的优势分析

传统时空预测方法在像素空间联合建模外观纹理与结构变化,预测误差在逐步外推中逐步放大,导致后期结果模糊。STA-Diff通过MRR显式建模跨阶段几何变换,将预测重点转移至运动表征空间,缓解外观变化对长期预测的干扰。实验表明,STA-Diff在LPIPS和SSIM指标上优势稳定,尤其在远距离预测阶段仍能保持连续的形态演化趋势,验证了运动与外观解耦建模在长跨度预测中的有效性。

3.2 临床意义

在IVF临床实践中,胚胎学家的决策高度依赖于对胚胎发育潜力的评估。传统的结果导向评估模式通常需要等待胚胎自然发育至囊胚期后,才能对其发育质量进行判断,这在一定程度上限制了移植决策的时间灵活性和调整空间。在此背景下,具备长跨度发育预测能力的生成模型为胚胎评估提供了一种新的可能性。本模型可在胚胎发育第3天提前预测第5-6天囊胚形态,为临床早期筛选、缩短培养周期、提高移植效率提供AI辅助依据。通过在较早发育阶段对后续形态进行预测,该方法有望为胚胎学家提供额外的参考信息,在不替代人工决策的前提下,辅助优化胚胎筛选与移植策略。此外,基于预测序列量化分析结果,也可能为构建更加客观和一致的胚胎评估标准提供支持。

3.3 局限性与未来工作

当前模型主要基于正常发育轨迹建模,对发育异常样本的预测能力有待提升。实验基于二维延时成像数据,未涉及三维结构信息及患者层面临床数据。未来工作将结合产妇年龄、激素水平等多模态临床信息进行联合建模,进一步提升预测的个体化准确性与临床适用性。

4 结论

本研究针对IVF胚胎发育长跨度预测过程中存在的非线性形变与误差累积问题,提出了一种基于扩散建模的预测框架STA-Diff。该方法通过在运动表征空间中进行条件扩散预测,有效缓解了直接在像素空间进行长序列外推所带来的结构不稳定性。

在方法设计上,本研究构建了一组协同作用的关键机制,包括标准化帧间残差先验(SIRP)、动态时空解耦门控混合器(D-STDm)以及时间加权条件调制(TWCM)。上述设计分别从结构引导、长程时空建模以及时间条件聚合等角度,对胚胎发育过程中的复杂演化特性进行了针对性建模。

实验结果表明,STA-Diff在多项图像质量指标以及关键结构区域一致性指标上均优于13种主流对比方法。尤其在长跨度预测场景中,当多种基线模型因误差累积出现性能退化甚至预测失效,失效模型数量由早期的2个增加至远期的9个时,STA-Diff仍能够保持稳定的预测表现。

本研究验证了在胚胎发育预测任务中,引入运动表征建模与条件扩散机制的可行性与有效

表 2 关键运动区域的定量对比

Table 2 Quantitative comparison on key motion regions

Model	MSE ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
ConvLSTM(Shi 等, 2015)	0.006476	21.89	0.3396	0.2164
E3D-LSTM(Wang 等, 2019)	0.005031	22.98	0.3326	0.2719
MIM(Wang 等, 2019)	0.005782	22.38	0.3428	0.2369
PhyDNet(Le 等, 2020)	0.007233	21.41	0.2816	0.1689
MAU(Chang 等, 2021)	0.006821	21.66	0.2776	0.0957
PredRNNv2(Wang 等, 2021)	0.005658	22.54	0.3578	0.193
SimVP(Gao 等, 2022)	0.005352	22.71	0.372	0.2431
TAU(Tan 等, 2023)	0.006388	21.95	0.3197	0.1681
MMVP(Zhong 等, 2023)	0.006751	21.71	0.3221	0.2836
SwinLSTM(Tang 等, 2023)	0.006278	22.02	0.3383	0.2137
WaST(Nie 等, 2024)	0.005485	22.61	0.3358	0.1741
RaMViD(Höppe 等, 2022)	0.009783	20.1	0.1706	0.1268
ExtDM(Zhang 等, 2024)	0.008186	20.87	0.3578	0.0728
STA-Diff	0.005098	22.93	0.3985	0.0534

注: 黑色字体表示最优结果。↓表示值越小越好; ↑表示值越大越好。

性。所提出的方法为胚胎发育过程的前瞻性分析提供了一种新的技术路径,并为后续探索更高维度时序数据建模及多模态临床信息融合奠定了基础。

致谢:本研究得到智能感知与计算安徽省高校重点实验室和安庆师范大学高性能计算平台的支持,在此表示衷心的感谢!

表 3 关键结构区域识别的定量对比
Table 3 Quantitative comparison of key structural region identification

Model	Dice ↑	IoU ↑	Ratio
ConvLSTM(Shi 等,2015)	0.4097	0.3024	0.3459
E3D-LSTM(Wang 等,2019)	0	0	0
MIM(Wang等,2019)	0.3114	0.2922	0.3044
PhyDNet(Le 等,2020)	0.2938	0.2626	0.2816
MAU(Chang 等,2021)	0.8634	0.7671	0.8367
PredRNNv2(Wang 等,2021)	0.3564	0.3042	0.3256
SimVP(Gao 等,2022)	0.1487	0.0957	0.0957
TAU(Tan 等,2023)	0.779	0.6452	0.7672
MMVP(Zhong 等,2023)	0	0	0
SwinLSTM(Tang 等,2023)	0.2404	0.188	0.2339
WaST(Nie 等,2024)	0.6123	0.5089	0.5743
RaMViD(Höppe 等,2022)	0.47865	0.3517	0.62335
ExtDM(Zhang 等,2024)	0.7263	0.6219	0.7341
STA-Diff	0.936	0.8798	1.0701

注:黑色字体表示最优结果;↑表示值越大越好。

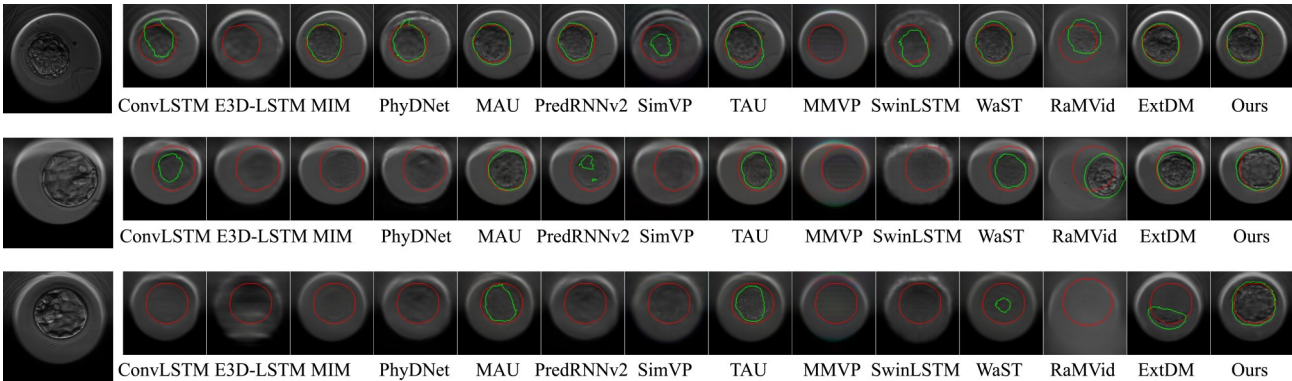


图6 使用Unet对囊胚图像关键运动区域标注真实标签和预测标签视觉对比

Fig. 6 Visual comparison of ground truth and predicted labels of key motion regions in blastocyst images using the U-Net model

表 4 消融实验结果

Table 4 Ablation study results

SIRP	D-STD	TWCM	MSE ↓	PSNR(dB) ↑	SSIM ↑	LPIPS ↓	Params (M) ↓	FLOPs (G) ↓
√	√	√	0.00412	23.85	0.6542	0.0571	64.5	1336
×	√	√	0.004483	23.48	0.6319	0.0695	64.5	1336
×	×	√	0.005176	22.86	0.6102	0.0781	51.5	1081
√	×	×	0.005842	22.31	0.5938	0.0847	51.07	1081
×	√	×	0.005461	22.57	0.6027	0.0816	64.07	1336
√	√	×	0.004931	23.02	0.6215	0.0739	64.07	1336
×	×	×	0.006521	21.86	0.6045	0.0932	51.07	1081

注:黑色字体表示最优结果。√表示使用对应模块;×表示不使用对应模块。↓表示值越小越好;↑表示值越大越好。

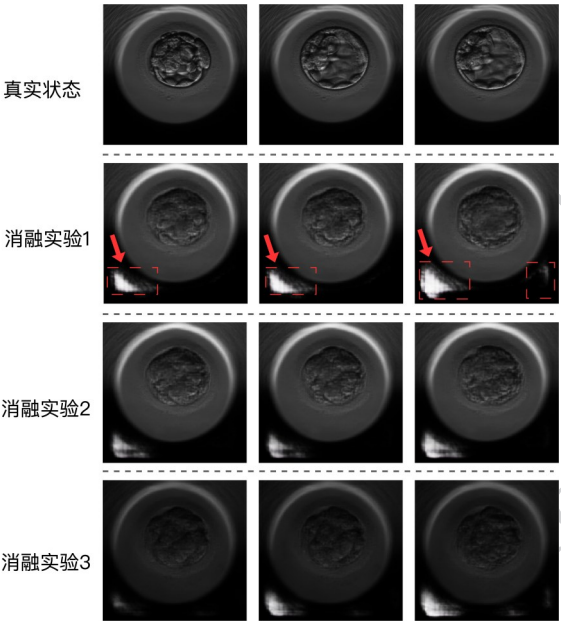


图 7 消融实验结果视觉对比
Fig. 7 Visual comparison of ablation study results

参考文献(References)

Araya B M, Velez M P, Gelaye K A, Dyer S and Aldersey H M. 2024. Addressing the rehabilitation needs of women experiencing infertility in Ethiopia: time for action. *International Journal of Environmental Research and Public Health*, 21(4): 475 [DOI: 10.3390/ijerph21040475]

Barriere P, Porcu-Buisson G and Hamamah S. 2018. Cost-effectiveness analysis of the gonadotropin treatments HP-hMG and rFSH for assisted reproductive technology in France: a Markov model analysis. *Applied Health Economics and Health Policy*, 16(1): 65-77 [DOI: 10.1007/s40258-017-0361-7]

Chang Z, Zhang X F, Wang S S, Ma S W, Ye Y, Xiang X G, et al.

2021. Mau: A motion-aware unit for video prediction and beyond// *Advances in Neural Information Processing Systems*, 34: 26950-26962

Dai C, Yang C L and Zheng Z B. 2019. Hierarchical multi-hypothesis prediction algorithm for compressed video sensing. *Journal of Image and Graphics*, 24(3): 357-365 (戴超, 杨春玲, 郑钊彪. 2019. 视频压缩感知中分级多假设预测算法. *中国图象图形学报*, 24(3): 357-365). [DOI: 10.11834/jig.180391]

Dirvanauskas D, Maskeliunas R, Raudonis V and Damasevicius R. 2019. Embryo development stage prediction algorithm for automated time lapse incubators. *Computer Methods and Programs in Biomedicine*, 177: 161-174 [DOI: 10.1016/j.cmpb.2019.05.027]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale [EB/OL]. [2026-04-27]. [DOI: 10.48550/arXiv.2010.11929]

Fruchter-Goldmeier Y, Kantor B, Ben-Meir A, Wainstock T, Erlich I, Levitas E, et al. 2023. An artificial intelligence algorithm for automated blastocyst morphometric parameters demonstrates a positive association with implantation potential. *Scientific Reports*, 13(1): 14617 [DOI: 10.1038/s41598-023-40923-x]

Gao Z Y, Tan C, Wu L R and Li S Z. 2022. Simvp: Simpler yet better video prediction//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE: 3170-3180 [DOI: 10.1109/CVPR52688.2022.00318]

Gomez T, Feyeux M, Boulant J, Normand N, David L, Paul-Gilloteaux P, et al. 2022. A time-lapse embryo dataset for morphokinetic parameter prediction. *Data in Brief*, 42: 108258 [DOI: 10.1016/j.dib.2022.108258]

Guo W, Liu S, Gong Z, Zhang G and Jiang X. 2023. Effective use of prior information for high-performance embryo grading. *Biomedical Signal Processing and Control*, 85: 104943 [DOI: 10.1016/j.bspc.2023.104943]

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic mod-

- els//Advances in Neural Information Processing Systems. Vancouver: Curran Associates, Inc.:6840-6851
- Höppe T, Mehrjou A, Bauer S, Nielsen D and Dittadi A. 2022. Diffusion models for video prediction and infilling [EB/OL]. [2026-04-27].
<https://arxiv.org/abs/2206.07696>
- Inhorn M C and Patrizio P. 2015. Infertility around the globe; new thinking on gender, reproductive technologies and global movements in the 21st century. *Human Reproduction Update*, 21 (4): 411-426 [DOI:10.1093/humupd/dmv016]
- Kragh M F, Rimestad J, Lassen J T, Berntsen J and Karstoft H. 2021. Predicting embryo viability based on self-supervised alignment of time-lapse videos. *IEEE Transactions on Medical Imaging*, 41 (2): 465-475 [DOI:10.1109/TMI.2021.3116986]
- Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc.:1097-1105 [DOI:10.1145/3065386]
- Le Guen V and Thome N. 2020. Disentangling physical dynamics from unknown factors for unsupervised video prediction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 11474-11484 [DOI: 10.1109/CVPR42600.2020.01149]
- Liu X J, Yu M X, Liu H H, Ma C L, Du W B, Wu H C, et al. 2025. DLT-Embryo: A dual-branch local feature fusion enhanced transformer for embryo multi-stage classification. *Biomedical Signal Processing and Control*, 102: 107266 [DOI: 10.1016/j.bspc.2024.107266]
- Lotter W, Kreiman G and Cox D. 2016. Deep predictive coding networks for video prediction and unsupervised learning [EB/OL]. [2026-04-27].
<https://arxiv.org/pdf/1605.08104.pdf>
- Ni H M, Shi C H, Li K, Huang S X and Min M R. 2023. Conditional image-to-video generation with latent flow diffusion models [EB/OL]. [2026-04-27].
<https://arxiv.org/abs/2303.13744>
- Nie X S, Yan Y F, Li S Y, Tan C, Chen X, Jin H Y, et al. 2024. Wavelet-driven spatiotemporal predictive learning: Bridging frequency and time variations//Proceedings of the 38th AAAI Conference on Artificial Intelligence: 4334-4342 [DOI: 10.1609/aaai.v38i5.28230]
- Ronneberger O, Fischer P and Brox T. 2015. U-net: Convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]
- Shi X J, Chen Z R and Wang H. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting [EB/OL]. [2026-04-27].
<https://arxiv.org/pdf/1506.04214.pdf>
- Tan C, Gao Z Y, Wu L R, Xu Y J, Xia J, Li S Y, et al. 2023. Temporal attention unit: Towards efficient spatiotemporal predictive learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 18770-18782 [DOI: 10.1109/CVPR54197.2023.01799]
- Tang S, Li C, Zhang P and Tang R N. 2023. SwinLSTM: Improving spatiotemporal prediction accuracy using Swin Transformer and LSTM [EB/OL]. [2026-04-27].
<https://arxiv.org/abs/2308.09891>
- Wang D Q and Zhao X. 2022. Class-aware network with global temporal relations for video action detection. *Journal of Image and Graphics*, 27 (12): 3566-3580 (王东祺, 赵旭. 2022. 类别敏感的全局时序关联视频动作检测. *中国图象图形学报*, 27 (12): 3566-3580) [DOI: 10.11834/jig.211096]
- Wang Z and Bovik A C. 2009. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26 (1): 98-117 [DOI:10.1109/MSP.2008.930649]
- Wang Y B, Jiang L, Yang M H, Li L J, Long M S and Li F F. 2019. Eidetic 3D LSTM: A model for video prediction and beyond [EB/OL]. [2026-04-27].
<https://openreview.net/pdf?id=B1IKS2AqtX>
- Wang Z, Bovik A C, Sheikh H R and Simoncelli E P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13 (4): 600-612 [DOI: 10.1109/TIP.2003.819861]
- Wang Y B, Zhang J J, Zhu H Y, Long M S, Wang J M and Yu P S. 2019. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE: 9154-9162 [DOI: 10.1109/CVPR.2019.00937]
- Wang Y B, Wu H X and Zhang J J. 2022. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (2): 2208-2225 [DOI:10.1109/TPAMI.2022.3165153]
- Wu C W, Yan W, Li H T, Li J X, Wang H K, Chang S J, et al. 2021. A classification system of day 3 human embryos using deep learning. *Biomedical Signal Processing and Control*, 70: 102943 [DOI: 10.1016/j.bspc.2021.102943]
- Yang L B, Peavey M, Kaskar K, Chappell N, Zhu L, Devlin D, et al. 2022. Development of a dynamic machine learning algorithm to predict clinical pregnancy and live birth rate with embryo morphokinetics. *F&S Reports*, 3 (2): 116-123 [DOI: 10.1016/j.xfre.2022.04.004]
- Zhang R, Isola P, Efros A A, Shechtman E and Wang O. 2018. The unreasonable effectiveness of deep features as a perceptual metric//Proceedings of the IEEE Conference on Computer Vision and Pat-

tern Recognition. Salt Lake City: IEEE: 586-595 [DOI: 10.1109/CVPR.2018.00068]

Zhang Y J, Zhang R Q, Zhou H J, Qi J, Yu Z F and Huang T J. 2025. Research status and development trends of vision foundation models. Journal of Image and Graphics, 30(1): 1-24 (张燚钧, 张润清, 周华健, 齐骥, 余肇飞, 黄铁军. 2025. 视觉基础模型研究现状与发展趋势. 中国图象图形学报, 30(1): 1-24) [DOI: 10.11834/jig.230911]

Zhang Z C, Hu J Y, Cheng W T, Paudel D and Yang J F. 2024. Extdm: Distribution extrapolation diffusion model for video prediction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 19310-19320

Zhao M, Xu M, Li H, Alqawasmeh O, Chung J P W, Li T C, et al. 2021. Application of convolutional neural network on early human embryo segmentation during in vitro fertilization. Journal of Cellular and Molecular Medicine, 25(5): 2633-2644 [DOI: 10.1111/jcmm.16288]

Zhong Y Q, Liang L, Zharkov I and Neumann U. 2023. MMVP: Motion-

Matrix-based Video Prediction//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris: IEEE: 4273-4283 [DOI: 10.1109/ICCV51070.2023.00394]

作者简介

陈义仁,男,副教授,主要研究方向为数字化骨科、医学图像智能分析等。E-mail:chenyr106@163.com

朱昌文,男,硕士研究生,主要研究方向为医学图像处理。E-mail:zcwyougemeng@163.com

赵家培,男,讲师,主要研究方向为大模型微调、计算机视觉等。E-mail:zhaojiapei2020@126.com

刘德阳,男,教授,主要研究方向为光场数据压缩与重建、智能视频编码等。E-mail:deyang.liu@hotmail.com

查红彬,男,教授,主要研究方向为智能科学技术、计算机视觉等。E-mail:zha@cis.pku.edu.cn

单彩峰,男,教授,主要研究方向为计算机视觉、模式识别、医学图像分析等。E-mail:cfshan@nju.edu.cn