第 14 卷 第 1 期
2025 年 1 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 14 No. 1
Jan. 2025

**引文格式：**
吴福祥, 程俊. 文本引导视频预测大模型的场景动态控制综述 [J]. 集成技术, 2025, 14(1): 9-24.
Wu FX, Cheng J. A review of scene dynamic control in text-guided video prediction large models [J]. Journal of Integration Technology, 2025, 14(1): 9-24.

# 文本引导视频预测大模型的场景动态控制综述

吴福祥 程 俊*

（中国科学院深圳先进技术研究院 深圳 518055）

**摘 要** 近年来，生成式人工智能的快速发展使文本驱动的视频预测大模型成为学术界和工业界的研究热点。视频预测生成需处理时间维度的动态性和一致性，要求精准控制场景结构、主体行为、相机运动和语义表达。当前的主要挑战是如何精确控制视频预测中的场景动态，以实现高质量和语义一致的输出。针对此问题，一些研究者提出了相机控制增强、参考视频控制、语义一致性增强和主体特征控制增强等方法，旨在提升视频预测的生成质量，确保生成内容既符合历史条件，又满足用户需求。该文系统探讨了上述 4 个控制方法的核心思想、优缺点和未来发展方向。

**关键词** 文本驱动视频预测；动态控制；相机控制；语义增强；主体特征控制
**中图分类号** TP391.7 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20241201002

# A Review of Scene Dynamic Control in Text-Guided Video Prediction Large Models

WU Fuxiang CHENG Jun*

( *Shenzhen Institute of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China* )

*Corresponding Author: *jun.cheng@siat.ac.cn*

**Abstract** In recent years, the rapid development of generative AI has made text-driven video prediction large models a hot topic in academia and industry. Video prediction and generation should address temporal dynamics and consistency, requiring precise control of scene structures, subject behaviors, camera movements, and semantic expressions. One major challenge is accurately controlling scene dynamics in video prediction to achieve high-quality, semantically consistent outputs. Researchers have proposed key control methods, including camera control enhancement, reference video control, semantic consistency enhancement, and subject feature control improvement. These methods aim to improve generation quality, ensuring outputs align with historical

**收稿日期**：2024-12-01 **修回日期**：2024-12-08
**基金项目**：国家自然科学基金项目 (U21A20487, 62372440)
**作者简介**：吴福祥，博士，副研究员，研究方向为多模态深度学习、图像生成；程俊(通讯作者)，博士，研究员，研究方向为机器视觉、智能机器人、人工智能，E-mail：jun.cheng@siat.ac.cn。

context while meeting user needs. This paper systematically explores the core concepts, advantages, limitations, and future directions of these four control approaches.

**Keywords** text-driven video prediction; dynamic control; camera control; semantic enhancement; subject feature control

**Funding** This work is supported by National Natural Science Foundation of China (U21A20487, 62372440)

## 1 引 言

近年来，生成式人工智能的快速发展使文本驱动的视频预测(text-guided video prediction，TVP)成为学术界和工业界的热点研究方向[1-5]。与静态图像生成相比，视频预测不仅需要捕捉每一帧的视觉信息，还必须处理时间维度上的连续性和动态性，因此描述的世界演变过程更丰富，如天空云朵演变、复杂海岸的逼真海浪和人类复杂运动等。TVP 根据给定的自然语言描述和历史视频片段生成后续的视频片段，在机器人示教、娱乐教育、世界模型[6-11]等领域的应用前景较广。由于文本能从不同层面描述世界，且文本，特别是自然语言，可以方便快捷构造，因此文本条件可以很好地描述给定历史视频的后续发展。然而，由于现实世界场景包含大量且多样的长尾物体及它们间的复杂交互，视觉模态往往以直接和详尽方式展现细节，而文本模态则包含抽象语义，文本模态和视觉模型的模态差异大，因此目前的文本引导的视频预测模型在视频相机控制、参考视频控制、文本语义一致性和视觉主体特征一致性等方面存在问题，这是场景动态控制不足导致的。为解决 TVP 中的控制问题，一些研究人员提出了不同层次的增强控制方法，本文主要讨论以下几个关键控制方法。

(1)相机控制增强：相机在视频预测生成中的运动轨迹和视角变化对展示场景的完整性和动态性至关重要。通过调整相机的运动路径、焦距变化和视角，生成系统能展示不同角度的场景，从而满足各种下游应用需求。

(2)参考视频控制：由于很多动作和运动比较少见，因此视频预测难以构建较少见的动作和运动视频。通过将较少见的动作和运动视频作为条件，与生成过程结合，控制生成内容迁移动作特征，可实现动作的匹配，并可有效提高生成效果。

(3)语义一致性增强：通过对输入文本进行深层次的语义分析，视频预测生成模型能更准确地识别重要的对象、动作和场景，并在视频预测生成中体现出来。这种语义增强方法能使生成的视频更符合文本描述。

(4)主体特征控制增强：主体控制主要聚焦于动态视频中关键对象或主体特征的生成与操控，主体特征表现是视频预测生成的关键，通过对视频中的人物、物体或场景的细节进行控制，可确保生成的主体符合描述，并可在时间维度上保持主体特征的一致性。

综上所述，TVP 中的动态控制既是提升生成视频质量的关键，又是确保生成内容满足用户特定需求的重要途径。场景动态控制是一个充满挑战的研究领域，通过不断完善相机控制增强、参考视频控制、语义一致性增强和主体特征控制增强等方法，可使文本视频预测更灵活，世界描述能力更强。本综述首先讨论视频预测的场景动态控制；然后，从相机控制增强、参考视频控制、语义一致性增强、主体特征控制增强等方面对场

景动态增强控制进行系统性探讨，并分析现有方法的核心思想、优缺点和未来发展方向。

## 2　文本引导视频预测

如图 1 所示，文本引导视频预测根据输入的文本描述预测与给定初始条件匹配的视频，是视频生成任务之一。目前有很多效果不错的开源和闭源模型，如 SORA[2]、Qingying、Hailuo、Ali Tongyi 和 CogVideoX[12-13]等❶。此类视频预测生成方法基本在潜特征空间里进行视频生成，可在降低运算量的同时提高生成质量。

### 2.1　视频潜特征

潜特征生成[14]不仅能在保持信息完整性的前提下大大降低存储和传输需求，还能提升后续处理(如生成、编辑等)效率。这个过程的关键在于将原始的高维视频数据压缩为低维度的空间表示，即“潜在空间”。编解码模型是实现上述转换的核心工具，通过变分自编码器或其改进版本工作。变分自编码器将输入数据映射到一个潜在空间，并学习该空间的有效概率分布。为进一步提高重建质量和模型的表达能力，研究者开始尝试结合其他技术，如对抗生成网络中的鉴别器损失或使用 Transformer 架构[15-18]处理图像和视频数据。

视频内容的编解码任务更复杂，不仅需要有效压缩空间维度信息，还需要捕捉时序信息。因此，最初的潜特征构造是尝试将视频中每帧视为独立的图像，通过图像编码器逐个进行编码，如在 Stable Video Diffusion[19]、Latent Shift[20]、VideoLDM[21]、Emu Video[22]和 CogVideoX[12-13]模型中。这种方法虽然直观易实现，但缺乏有效的跨时间压缩，难以处理长时间序列或高帧率视频内容。为克服这些问题，可以在时间维度引入池化层降采样，如 MAGVIT[23]模型，也可以像 W.A.L.T.[24]和 MAGVIT-V2[25]模型使用 3D 卷积操作编码整个视频片段。由于 3D 卷积操作的计算量较大，因此，Movie Gen[26]使用交错 2D 和 1D 卷积层(2.5D 混合架构)处理时间和空间维度的信息，以保持低计算成本。

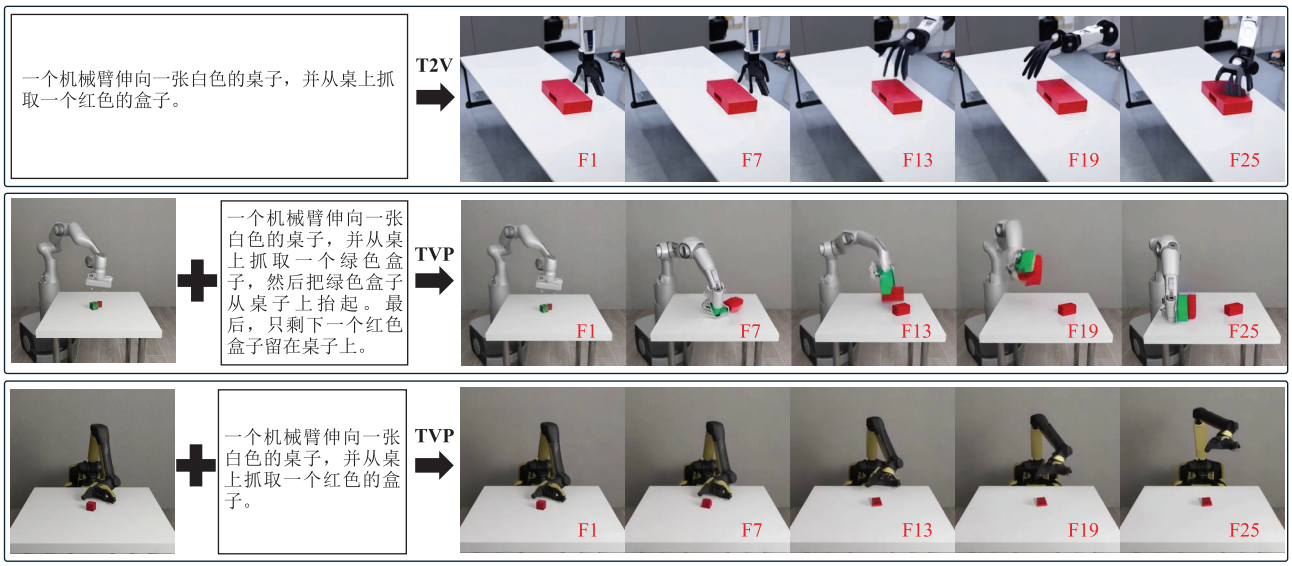


**图 1　文本生成视频与文本引导视频预测**

**Fig. 1　Text-to-video and text-guided video prediction**

注❶：Qingying：https://chatglm.cn；Hailuo：https://hailuoai.com/video；Ali Tongyi：https://tongyi.aliyun.com/wanxiang；CogVideoX-Fun 5B：https://github.com/aigc-apps/CogVideoX-Fun。

### 2.2 文本引导的视频生成

VideoCrafter2[27]探索了从 Stable Diffusion 扩展的视频模型训练方案，利用低质量视频和合成的高质量图像构建高质量视频模型。通过解耦前景物体运动与外观，该方法可在保证运动一致性的同时提升视频的画质和概念构成，从而提升文本到视频生成的质量。Text2Video-Zero[28]基于现有文本到图像合成方法，如 Stable Diffusion，提出一种零样本文本到视频生成的新任务，实现了低成本、高质量、一致性强的视频生成，创新点在于通过引入运动相关动态特征改进生成帧的潜在编码和跨帧注意力机制。TF-T2V[29]构造了 3D-UNET 预测噪声，通过分离文本解码、时间建模的过程，以及内容分支和运动分支进行联合优化。由于公开数据集的规模有限，因此该方法可仅使用无标签视频进行学习，也可以重新引入部分文本标签协同训练。HiGen[30]基于 3D-UNET 从结构层面和内容层面对视频的空间和时间因素进行解耦。在结构层面，它将文本到视频的任务分为空间推理和时间推理，利用统一去噪器生成空间一致的先验和时间一致的运动。在内容层面，通过提取输入视频中的运动和外观变化线索，指导模型训练和生成视频，增强了内容的灵活性和时间稳定性。Snap Video[31]通过扩展解析扩散生成框架适应空间和时间冗余像素，同时一种新型的基于 Transformer 的架构[32-33]加快了生成模型的训练速度和推理速度，显著提高了视频质量、时间一致性和运动复杂性。PDMs（patch diffusion models）[34]建模视频分块分布，而非整个输入，可极大程度地提高高分辨率视频的训练效率。PDMs 通过深度上下文融合，确保视频分块间的一致性。此外，PDMs 可通过微调将低分辨率生成器快速转换为高分辨率文本到视频合成模型。GLOBER[35]提出了非自回归的视频生成方法，可生成全局特征，进而对生成进行全局指导，然后基于全局特征合成视频帧，以非自回归方式生成连贯的视频，该方法先通过视频编码器提取全局特征，再基于扩散模型来合成视频帧，允许灵活合成任意子视频，并采用新型对抗损失提高全局一致性和局部真实感。

### 2.3 文本引导的视频预测

基于视频生成技术，TVP 可通过文本条件，从一张或多张历史帧中预测场景的未来变化。如果仅使用一张历史帧进行预测，则构成更复杂的任务，即文本引导的图像生成视频（T-I2V）。通过给定参考图像生成视频在角色动画生成等方面的应用价值非常重要。Animate Anyone[36]通过引入 ReferenceNet，并通过空间注意机制融合参考图像的细节特征，以保持复杂外观特征的一致性。该方法还构造了一个姿势引导器，能有效控制角色的运动，并能确保角色动作的连续性，通过采用有效的时序建模方法，确保视频帧之间的平滑过渡。Motion-I2V[37]提出两阶段的文本指导图像到视频预测生成。第一阶段提出一种基于扩散的运动场预测器，专注于推导参考图像像素的轨迹。第二阶段引入了运动增强的时间注意机制，增强了视频潜在扩散模型中有限的一维时间注意力，能有效地将参考图像特征传播到合成帧中，并依据第一阶段预测的轨迹进行引导，在面对大运动和视角变化时生成更一致的视频。此外，通过为第一阶段训练稀疏轨迹的 ControlNet，可支持用户精确控制运动轨迹和区域，提供比单纯依赖文本指令更高的可控性。TI2V-Zero[38]旨在实现基于给定图像和文本描述的真实视频合成，通过预训练的文本到视频扩散模型，采用“重复滑动”策略调节去噪过程，逐帧合成视频，并使用 DDPM（denoising diffusion probabilistic models）反演策略确保时序连续性。TI2V-Zero 可扩展到视频填充和预测等其他任务，支持长视频预测生成。Still-Moving[39]通过对定制的文本到图像模型进行调整，提高生成效果。该方法提出了轻量级空间适配器，可调整从

Text to Image (T2I)层产生的特征，解决了文本到图像模型权重直接嵌入至文本到视频模型所带来的伪影和不符合文本条件的问题。该方法还引入了新颖的运动适配器模块，允许保持视频的运动先验。DriveDreamer-2[11]通过统一的多视图模型，能够生成遵循大语言模型约束的高清地图，以及符合交通法规的多视角驾驶视频，能通过给定初始帧预测后续的驾驶情景。GR-2[40]是一个通用机器人代理，通过在大规模互联网视频片段上进行预训练和在多种机器人任务上进行微调构建，能基于一些初始的历史帧预测后续的视频，可为机械臂提供示教视频。

与文本生成视频相比，TVP 通过引入初始条件图像或视频使生成的视频更符合历史条件，从而更好地控制场景的动态变化。然而，与文本生成视频类似，这种方法仍然面临相机控制、复杂动作和外观一致性问题，以及细粒度外观难以控制等问题，需进一步结合相关条件实现更精细的控制。

### 2.4　基准测试和评价指标

视频生成中，常用的指标包括 Fréchet Inception Distance (FID)[41]、Fréchet Video Distance (FVD)[42]和 CLIP[43]等，应用于 PDMs[34]、HiGen[30]、Text2Video-Zero[28]、TF-T2V[29]、Snap Video[31]和 GLOBER[35]等。其中：HiGen[30]还使用了基于 CLIP 的 Temporal Consistency[44]，以评价帧间的相似性；而 PDMs[34]使用了 Inception Score (IS)[45]指标等。此外，VideoCrafter2[27]利用 EvalCrafter 进行评价，EvalCrafter 提供 512 个提示词，使用 18 个客观指标评价生成视频的视觉质量、内容质量和文本-视频对齐情况。此外，人类角度的用户偏好研究则通过要求 3 个视频制作专家从视频集中选择偏好视频完成。

在视频预测上，TI2V-Zero[38]提出基于文本条件的 FVD 和基于主题条件的 FVD，并利用 FVD[42]指标衡量生成视频与文本提示和给定图像的匹配程度。DriveDreamer-2[11]使用 FID 和 FVD 作为评估指标。Still-Moving[39]通过 CLIP 文本-图像相似度和每帧相对于条件图像的 CLIP 图像-图像相似度评估生成视频的质量。Animate Anyone[36]在图像级别的质量定量评估中采用 SSIM[46]、PSNR[47]和 LPIPS[48]等指标。ConsistI2V[49]使用包含主体身份一致性、背景一致性、时序闪烁和运动流畅性等多个指标的 VBench[50]评估视频质量，该方法还包括物体类别、多物体、人物动作和空间关系等视频质量评估指标。Motion-I2V[37]构建了一个涵盖多种类别的测试集，并使用了 80 张无版权图片，通过 ChatGPT-4V 为每张图片的内容和可能的动作生成提示词，利用 CLIP 文本-图像对数衡量提示词与生成帧的一致性。GR-2[40]将生成视频中的任务成功率作为评判标准。

视频预测的评估极具挑战，因为通用指标通常难以准确反映人类的感知和偏好，且无法全面衡量各种主题下的表现，所以通常利用 FVD、CLIP 和 VBench 等较通用的指标评估视频的质量和一致性，此外，还会引入特定领域的评价标准，如 GR-2 中的成功率，以确保评估的完整性。

## 3　视频预测中的场景动态增强控制

场景动态控制需要处理时间和空间维度上的动态性，使预测生成模型更好地跟随多模态条件预测场景世界的演变。如表 1 所示，常采用相机控制增强、参考视频控制、语义一致性增强与主体特征控制增强等完成视频场景的增强控制。

### 3.1　视频预测中的相机控制增强

相机控制具有非常重要的作用，相机的移动会加大场景中各物体运行的复杂度，如模拟安装在机械臂上的相机，其示教视频的生成需较好地模拟相机的运动和该相机视角下的物体

表 1 视频预测生成中的场景动态控制增强相关方法

Table 1 Enhancing scene dynamic control and related methods for video prediction generation

| 动态控制 | 控制生成方法 |
|---|---|
| 相机控制增强 | Free3D[51]、CamCo[52]、CameraCtrl[53]、Text-Animator[54]、FlowZero[55]、VideoStudio [56]、ConsistI2V[49]、MotionBooth[57]、AnimateDiff[58]、Movie Gen[26]、SEINE[59]、Direct-a-Video[60]、MotionCtrl[61] |
| 参考视频控制 | DreamVideo[5]、MotionClone[62]、SV4D[63]、CustomCrafter[64] |
| 语义一致性增强 | DirecT2V[65]、FlowZero[55]、LaDi[66]、VideoStudio[56]、AID[67]、ViD-GPT[68]、GPT4Motion[4]、PhysGen[69] |
| 主体特征控制增强 | Magic-Me[70]、ID-Animator[71]、VideoSwap[72]、DisenStudio[73]、DreamVideo[5]、MotionBooth[57] |

交互。目前的 TVP 模型很难直接控制相机，如图 2 所示。为精确控制相机，目前的不少方法借助 Plücker 坐标编码相机位置和运动，如 Free3D[51]、CamCo[52]、CameraCtrl[53]和 Text-Animator[54]。其中，Free3D[51]从预训练的 2D 图像生成器出发，通过微调实现新视角合成，通过使用 Plücker 坐标，引入新的光线条件归一化层，可将姿态信息注入底层 2D 图像生成器，通过轻量级的多视角注意力层和共享不同视角之间的生成噪声，可进一步增强多视角一致性；CamCo[52]通过使用 Plücker 坐标，为预训练的图像到视频生成器提供精确参数化的相机姿态输入，通过在每个注意力模块中整合极点注意力模块，增强生成视频中的 3D 一致性，确保特征图符合极点约束，通过对实际视频进行微调，并通过运动结构算法估计相机姿态，提升物体运动的合成效果，该方法使用户能更精确地控制视频生成中的相机位姿；CameraCtrl[53]使用 Plücker 坐标精确参数化相机轨迹，并在 T2V 模型上训练一个即插即用的相机模块，其他部分保持不变，此外，CameraCtrl 探讨了不同数据集对可控性和泛化能力的影响，结果表明，具有多样相机分

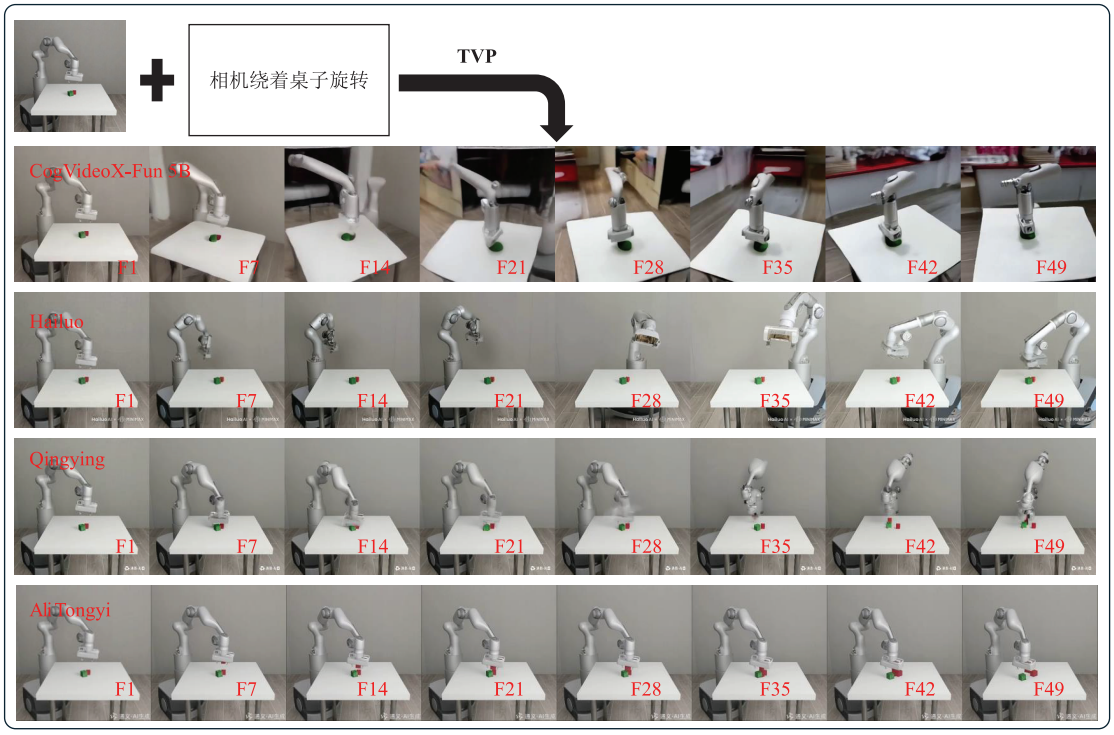


图 2 文本引导视频预测生成中的相机控制

Fig. 2 Camera control in text-guided video prediction generation

布和相似外观的视频的可控性和泛化性被提升；Text-Animator[54]通过文本嵌入注入模块精确描绘生成视频中的视觉文本结构，并结合基于类似 Plücker 坐标的相机控制模块和文本细化模块提升生成视觉文本的稳定性，旨在应对文本在生成视频中有效可视化的挑战，这在游戏、电商和广告等多个行业中至关重要。这些方法都利用 Plücker 坐标精确编码相机位置和运动，分别在新视角合成、视频生成和文本到视频生成中实现了更好的 3D 一致性和用户控制，推动了动态视频叙事的发展。此外，为提供更精细的运动控制，MotionCtrl[61]通过相机姿势和轨迹控制视频生成。它提出一个统一、灵活的运动控制器，能有效、独立地控制视频生成中的摄像机运动和物体运动，且其运动控制条件由摄像机的位姿和轨迹决定，不依赖物体的外观或形状，在生成视频时，对物体的外观和形状影响最小，这种设计使得模型能更专注运动控制，而非外观改变，提升了视频生成的质量。

Plücker 坐标虽然可以精确编码相机的位置和运动，但由于用户也常用如放大、缩小等方式概要描述相机运动，也会指定不同类型镜头，需要支持较为抽象控制条件[26]，因此 ConsistI2V[49]、MotionBooth[57]、AnimateDiff[58]、Movie Gen[26]、SEINE[59]和 Direct-a-Video[60]等方法基于概要相机运动描述控制视频生成。Ren 等[49]提出了基于扩散的方法 ConsistI2V，可增强文本生成视频的视觉一致性。该方法通过引入时空注意力机制，在初始帧上进行时空注意，可确保空间和运动的一致性，并可实现摄像机摇摄和放大/缩小效果。该方法从初始帧的低频带进行噪声初始化，可增强布局一致性。MotionBooth[57]通过少量对象图像对文本生成视频模型进行高效微调，能准确捕捉对象的形状和属性，并能精确控制对象和摄像机的运动。此外，Wu 等[57]在推理阶段提出了无须训练的对象和摄像机运动控制方法，通过跨注意力图操作控制对象运动，并通过引入新的潜在偏移模块控制摄像机运动。AnimateDiff[58]提出一个即插即用的运动模块，该模块能通过一次性训练后无缝整合到任何同一基础文本到图像（T2I）生成的个性化 T2I 中，还能有效学习来自真实视频的可转移运动先验，一旦训练完成，即可嵌入个性化 T2I 模型，形成个性化动画生成器。此外，AnimateDiff[58]还提出了 MotionLoRA，一种轻量级微调技术，可使预训练的运动模块以较低的训练和数据收集成本适应新运动模式，如相机变焦、摇摄和旋转等。Movie Gen[26]通过构造相机运动分类器预测放大、缩小、推进、拉远、摇摄等 16 种相机运动类型，并利用这些相机运动控制训练 Transformer 生成模型，实现电影级的相机运动控制，同时在微调中兼顾了广角、特写和航拍视角等 6 种额外镜头类型。SEINE[59]通过构造随机遮罩视频扩散模型实现包含相机放大与缩小的连贯长视频生成，在扩散模型中，基于文本描述自动生成场景过渡视频，可确保视觉质量和连贯性，通过输入不同场景的图像和文本控制，可生成平滑、创意性的过渡，适应不同长度的镜头级视频。Direct-a-Video[60]针对现有模型在控制物体运动和相机运动上无法独立操作的问题，解耦了对象运动和相机运动。该方法利用空间交叉注意力调制技术和模型的固有先验，无须额外的优化过程即可控制物体运动。此外，该方法引入了一种新的时间交叉注意力层，可解释平移和缩放等相机运动。

由于场景具有高动态性，因此 FlowZero[55]和 VideoStudio[56]通过结合大语言模型（large language model，LLM）生成高质量动态场景视频。其中，FlowZero[55]通过结合 LLM 和图像扩散模型，并通过生成动态场景语法来理解复杂的时空动态，指导图像扩散模型生成具有平滑运动和帧间一致性的视频，提升相机运动的真实性，同时通过迭代自我优化过程提升时空布局与文本

提示之间的对齐，并通过在每帧的初始噪声中加入运动动态，增强整体一致性，实现更生动的零样本动态多场景视频合成；VideoStudio[56]通过结合 LLM 控制相机运动信息，生成内容一致的多场景视频。它将输入提示转换为综合的多场景剧本，成功管理了场景间的逻辑，同时保持关键内容的一致视觉效果。每个场景的剧本包括事件描述、前景和背景实体，以及相机运动信息。通过识别剧本中的共同实体，并通过请求 LLM 对每个实体进行详细描述，可使这些描述用于生成参考图像。最终，通过扩散过程生成每个场景的视频，参考图像和事件描述被用作条件，增强了多场景视频的内容一致性。

在视频预测生成中，相机控制虽然已经取得了显著进步，但由于现实世界的复杂性，以及相机运动需要解决前景与背景运动的协调性，且相机位置不同会导致预测生成模型需要推理新视角下的场景，因此加大了视频预测生成的复杂性。总之，相机控制仍需解决前景与背景视觉一致性、运动控制精细度、场景主体的一致性、多视角泛化能力等方面的问题。可以考虑通过引入多层次运动建模与协同优化策略，结合前景背景分离自监督学习和跨视角对比生成，提升相机控制生成的表现。

### 3.2 视频预测中时序上的参考视频控制

考虑到复杂场景的多样性以及动作交互的复杂性，许多物体的时序交互和动作呈长尾分布，导致视频预测在构建此类视频时异常困难。此外，由于文本描述具有抽象性，难以涵盖场景中的细节，因此进一步增加了预测难度，如图 3 所示。而参考视频可以补充此类时序动作等信息，例如，DreamVideo[5]等通过迁移参考视频动作对动作生成进行控制。

给定目标动作视频，DreamVideo[5]能通过动作学习生成个性化的动作视频。该动作学习通过设计运动适配器，并基于给定视频有效地建模目标动作模式。该运动适配器能基于一类视频、多个展现相同动作的视频或单个视频中提取的动作模式进行动作定制。MotionClone[62]可通过参考视频实现运动克隆，进而控制文本到视频的

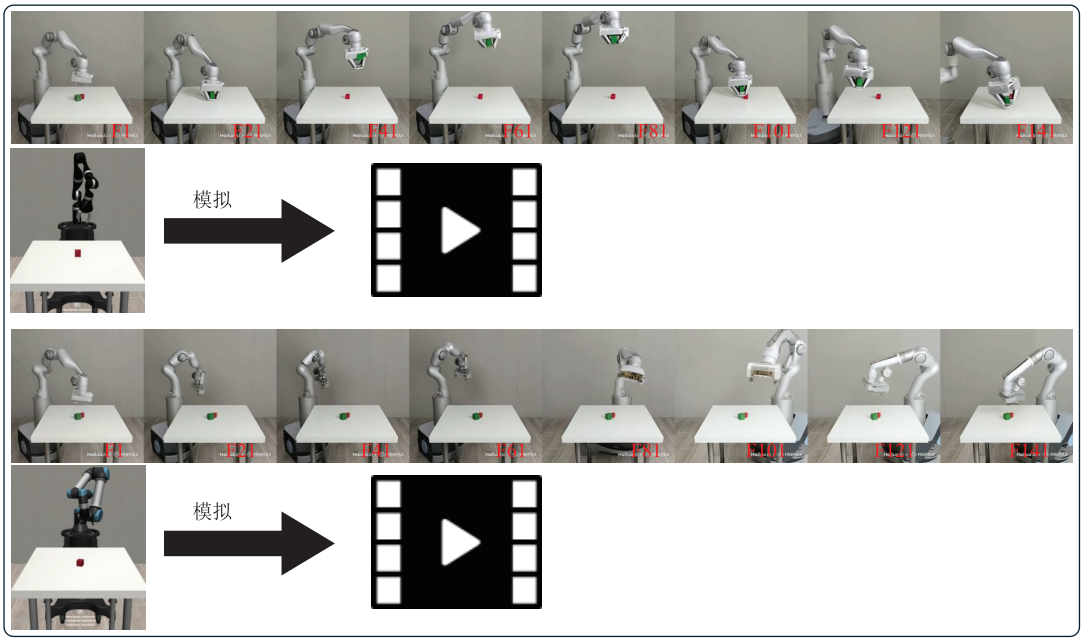


图 3 视频预测生成中的视频参考生成问题

**Fig. 3 Video reference generation problem in video prediction generation**

生成。它采用时间注意力机制控制参考视频中的运动，并引入初级时间注意力引导，以减轻噪声或细微运动对生成的影响，同时通过位置感知语义引导机制，结合参考视频中的前景粗略位置和无分类器引导特征，优化空间关系和文本遵循能力。SV4D[63]设计了旨在生成多帧和多视角一致的统一扩散模型，能从单目参考视频中生成时间一致的新视角视频，并通过动态 NeRF 高效优化隐式 4D 表示，避免了传统方法中繁琐的优化过程。CustomCrafter[64]设计了一个轻量化可插拔模块，通过融入视频扩散模型提升模型对新主题的外观细节捕捉和概念组合能力。通过观察可知，视频扩散模型在去噪初期更关注视频运动，在去噪后期更关注细节恢复，提出了动态加权视频采样策略。其在去噪初期通过小权重减少该可拔插模块对运动生成的影响，去噪后期再启用该模块，以修复指定主题的外观细节，确保外观的真实感。

参考视频虽然能较好地引导视频生成，但为确保模型能提取有效的视觉特征，基于参考视频预测生成视频高度依赖高质量数据。此外，生成视频的细节控制方面仍存在挑战，例如，生成视频难以高质量地还原参考视频中的精细动作或细粒度纹理。因此，降低对高质量数据的依赖并增强对视频细节的控制能力，成为提升参考视频控制效果的关键环节。通过引入多尺度学习增强模型对低质量数据的适应性，并通过结合自监督学习和细粒度特征优化，可有效降低该预测生成任务对高质量数据的依赖，并可提升视频细节控制能力。

### 3.3　视频预测中的语义一致性增强

如图 4 所示，由于文本与视频模态之间存在较大差异，因此文本对视频的预测难以实现细粒度控制，可能原因是文本描述不够具体，从而导致生成模型难以准确理解复杂场景中的时空动态。为解决这一问题，可结合 LLM 和文本增强技术等，补充不同层次的描述信息，从而有效提升视频生成的连贯性、一致性和质量，最终改善复杂动态场景中的视频预测效果。

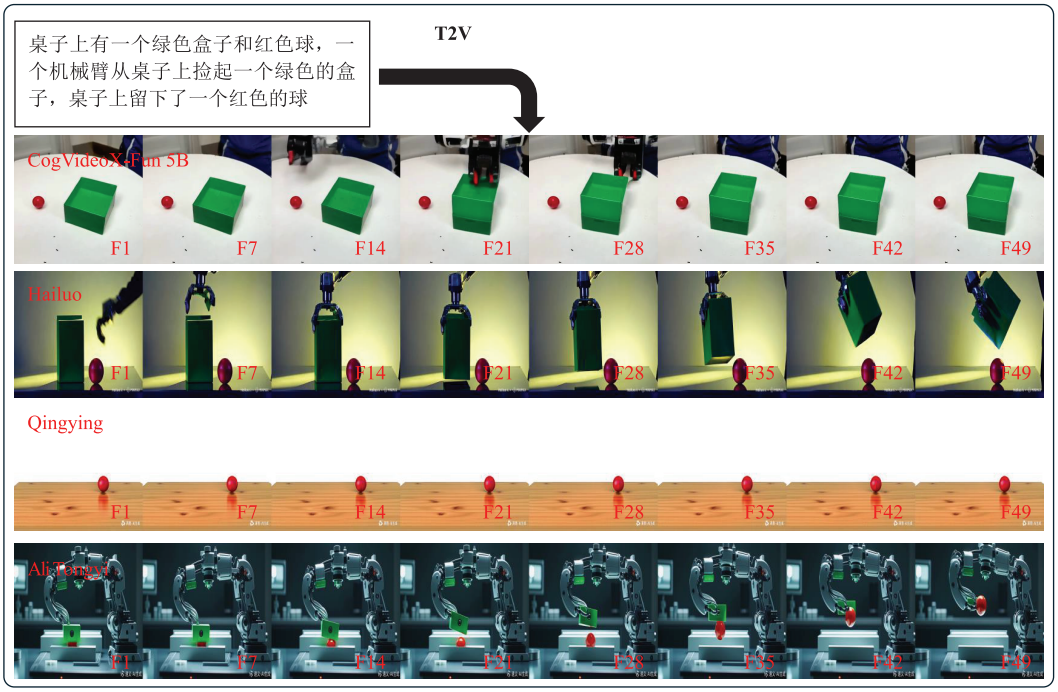


图 4　视频预测生成中的语义一致性问题

**Fig. 4　Semantic consistency issue in video prediction generation**

通过增强文本描述，可更好地控制视觉媒体的生成[74-81]，通过增强对文本情境知识的理解，能更有效地构建逻辑性强且连贯的场景及角色互动。DirecT2V[65]基于预训练的文本到图像（T2I）模型，借助 LLM 生成时序和逐帧提示，并将经过指令调优的 LLM 作为导演。该模型能生成包含时间变化的内容，并能促进视频生成的一致性。此外，为保持时间一致性，并防止将值映射到不同对象，DirecT2V 为扩散模型配备了一种不需要额外训练的值映射方法。FlowZero[55]利用 LLM 理解复杂的时空动态，生成动态场景语法，涵盖场景描述、物体布局和背景运动模式，以指导图像扩散模型生成平滑的物体运动和帧间一致性。此外，FlowZero 利用迭代自我优化过程提高时空布局与文本提示之间的对齐。通过增强每帧的初始噪声和融入运动动态，增强全局一致性，可实现零样本视频合成。由于文本到图像生成的有效性常受到文本提示不足的限制，导致生成的图像在艺术上的一致性和主题相关性不足，因此，为提升图像和视频生成的质量，LaDi[66]将 LLM 作为艺术指导。通过在约束解码、智能提示、微调方面整合增强文本到图像和文本到视频生成器，可提升视频生成质量。VideoStudio[56]利用 LLM 将输入提示转换为多场景脚本。通过借助 LLM 的逻辑知识，该生成的脚本包含场景事件描述、前景/背景实体和镜头运动等。同时 VideoStudio 识别脚本中的共同实体利用 LLM 细化每个实体描述，然后通过文本到图像模型生成实体的参考图像。而后，扩散模型通过参考图像、事件描述和镜头运动生成各个场景视频，增强了多场景视频的内容一致性。

利用多模态大语言模型，可以进一步控制视频生成。在根据指令和初始帧预测未来帧时，由于存在帧一致性和时间稳定性不足的问题，因此 AID[67]基于图像到视频 Image2Video 扩散模型，引入多模态大语言模型，用于根据初始帧和文本指令预测未来视频状态。通过双查询 Transformer，将指令和帧整合到条件嵌入中，可预测未来的帧。此外，通过构造长短期时间适配器和空间适配器，可以以最小的训练成本快速将通用视频扩散模型转移到特定场景中。由于生成时间一致性长视频较难，因此 ViD-GPT[68]通过将过去的帧作为“提示”生成未来的帧，让每一帧仅依赖之前的帧以满足因果生成约束，可实现单向生成。此外，通过在时间轴上将条件帧与待生成帧拼接，强化长程上下文获取，可合成较长视频。GPT4Motion[4]利用大语言模型 GPT-4 的规划能力生成 Blender 脚本，并通过物理引擎创建连贯的物理运动场景，然后控制文本到图像扩散模型，无须额外训练生成视频。PhysGen[69]提出结合基于模型的物理模拟和数据驱动的视频生成过程，能通过单个图像和输入条件（如施加在图像物体上的力和扭矩）生成真实、物理合理且时间一致的视频。PhysGen 包含捕捉图像的几何、材质和物理参数的多模态大语言模型（MLLM）图像理解模块，并通过刚体物理模拟器模拟真实行为，最后通过视频扩散模型生成逼真动画。

通过引入大语言模型补充文本信息虽然能提升生成效果，但其计算量较高，限制了实际应用。此外，如何更高效地利用大语言模型中隐含的知识，从多个信息层级补充场景中物体运动的相关常识，进一步规范视频生成，以符合现实规律，是未来的重要方向。通过引入轻量化的知识蒸馏和多层次的上下文建模，结合运动规律与场景约束，能在减少计算量的同时提升视频生成的真实性和达到符合物理常识的效果。

### 3.4 视频预测中的主体特征控制

如图 5 所示，视频预测生成中的场景主体常常会变形和扭曲。因此，复杂场景下的视频生成面临诸多挑战，如保持主体外观一致性等。针对上述问题，可通过主体控制提升生成视频中主体的适宜性，主要涉及身份生成、面部替换和多主

体定制等关键方面。

在主体外观控制方面，Magic-Me[70]通过微调生成特定身份信息的视频。而 ID-Animator[71]利用面部适配器替代视频中的人物主体。Magic-Me[70]是一种生成指定身份的视频框架，通过使用少量参考图像在初始化阶段注入帧间关联，并通过引入一种新的噪声初始化方法增强帧间稳定性。最后，Magic-Me 使用基于扩展文本反演训练的身份模块和裁剪后的身份图像分离身份信息和背景，在提高分辨率的同时强化面部特征与保持身份特征。ID-Animator[71]无须进一步训练，利用扩散视频生成架构和人脸适配器编码身份相关特征，可根据单一面部图像生成个性化视频。VideoSwap[72]提出定制的视频主体替换，旨在将源视频中的主要主体替换为具有不同身份和可能不同形状的目标主体。VideoSwap 利用语义点对应关系，借助少量语义点即可对齐主体的运动轨迹，并改变主体形状。

DisenStudio[73]支持生成定制化的多主体视频，通过引入空间解耦交叉[2]机制，增强预训练的扩散式文本到视频模型，使每个主体与其所需动作相结合，解决了现有方法在多主体场景中的缺失和绑定问题。此外，通过多主体共现、单主体掩蔽和多主体运动保留微调，该方法可保留主体和视觉属性，同时在静态图像上微调时保持时间运动生成能力。在给定目标主体图像和主体控制文本时，DreamVideo[5]通过主体学习实现定制，主体学习利用文本反演和身份适配器进行细致的外观捕捉，从而准确捕捉所提供图像中的主体细节特征。MotionBooth[57]使用特定对象的少量图像，有效地微调了文本到视频模型，以准确捕捉对象的形状和特征，从而控制视频的主体。

主体控制有助于提升预测视频的质量和语义一致性，但仍面临多主体交互场景下解耦控制的挑战，以及复杂动作的泛化能力和生成视频的逼真度与自然度不足问题。此外，主体的细粒度特征难以精确控制。通过引入多尺度控制、强化学习和自监督学习，优化主体间特征交互，提升细

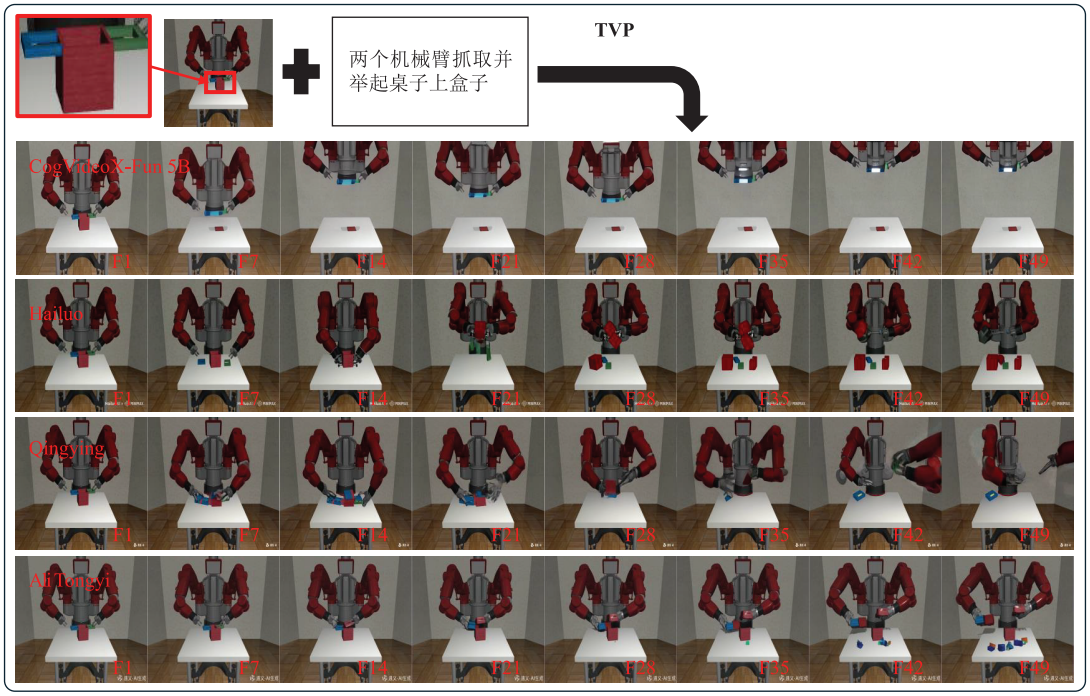


图 5　视频预测生成中的主体一致性问题

**Fig. 5　Subject consistency issue in video prediction generation**

粒度特征的表述质量与控制能力，可改善复杂场景下的泛化能力和提升视频生成的逼真度与自然度。

## 4 总 结

近年来，作为生成式人工智能领域的前沿研究方向，文本驱动的视频预测面临如何更有效地控制视频内容动态性和一致性的挑战。本文分析了视频预测的基本方法，并综述了 4 个关键的场景控制方法，旨在提高生成视频的质量和一致性。首先，通过相机控制实现运动轨迹和视角调整，以丰富场景表现力；其次，参考视频控制技术将现有的视觉素材作为参照，确保生成的内容在外观和动作上一致；再次，通过对输入文本进行深度分析，语义增强可准确识别关键对象和动作，进而提高视频内容的语义符合度；最后，主体控制着重于动态视频中关键对象的行为设定，使其与描述相匹配。通过系统探讨这 4 个方向，本文深入分析了当前技术的优势与不足，可为未来研究提供参考。总体而言，动态控制在提升视频预测质量及满足用户需求方面发挥重要作用。然而，由于现实场景的复杂性，目前的技术在处理大规模动态场景等挑战时仍显不足，因此未来的研究需进一步强化复杂动态场景生成能力和适应复杂动态情境等。

## 参 考 文 献

[1] Ma HY, Mahdizadehaghdam S, Wu BC, et al. MaskINT: video editing via interpolative non-autoregressive masked Transformers [C] // Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 7403-7412.

[2] Brooks T, Peebles B, Holmes C, et al. Video generation models as world simulators [R/OL]. (2024-02-15)[2024-12-01]. https://openai.com/research/video-generation-models-as-world-simulators.

[3] 邓梓焌, 何相腾, 彭宇新. 文本到视频生成: 研究现状、进展和挑战 [J]. 电子与信息学报, 2024, 46(5): 1632-1644.
Deng ZJ, He XT, Peng YX. Text-to-video generation: research status, progress and challenges [J]. Journal of Electronics & Information Technology, 2024, 46(5): 1632-1644.

[4] Lü JX, Huang Y, Yan MF, et al. GPT4Motion: scripting physical motions in text-to-video generation via Blender-oriented GPT planning [C] // Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2024: 1430-1440.

[5] Wei YJ, Zhang SW, Qing ZW, et al. Dream video: composing your dream videos with customized subject and motion [C] // Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 6537-6549.

[6] Alonso E, Jelley A, Micheli V, et al. Diffusion for world modeling: visual details matter in Atari [Z/OL]. arXiv Preprint, arXiv: 2405.12399, 2024.

[7] Meng FQ, Liao JQ, Tan XY, et al. Towards world simulator: crafting physical commonsense-based benchmark for video generation [Z/OL]. arXiv Preprint, arXiv: 2410.05363, 2024.

[8] Xiang JN, Liu GY, Gu Y, et al. Pandora: towards general world model with natural language actions and video states [Z/OL]. arXiv Preprint, arXiv: 2406.09455, 2024.

[9] Yang S, Du YL, Ghasemipour SKS, et al. Learning interactive real-world simulators [C] // The Twelfth International Conference on Learning Representations, 2024: 1-25.

[10] Yang S, Walker J, Parker-Holder J, et al. Video as the new language for real-world decision making [Z/OL]. arXiv Preprint, arXiv: 2402.17139, 2024.

[11] Zhao GS, Wang XF, Zhu Z, et al. DriveDreamer-2: LLM-enhanced world models for diverse driving video generation [Z/OL]. arXiv Preprint, arXiv: 2403.06845, 2024.

[12] Hong WY, Ding M, Zheng WD, et al. CogVideo: large-scale pretraining for text-to-video generation via Transformers [C] // Proceedings of the https://

openreview.net/forum?id=rB6TpjAuSRy.

[13] Yang ZY, Teng JY, Zheng WD, et al. CogVideoX: text-to-video diffusion models with an expert Transformer [Z/OL]. arXiv Preprint, arXiv: 2408.06072, 2024.

[14] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10684-10695.

[15] Cao SY, Yin YQ, Huang LH, et al. Efficient-VQGAN: towards high-resolution image generation with efficient vision Transformers [C] // Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision, 2023: 7368-7377.

[16] Chefer H, Alaluf Y, Vinker Y, et al. Attend-and-excite: attention-based semantic guidance for text-to-image diffusion models [J]. ACM Transactions on Graphics, 2023, 42(4): 1-10.

[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] // Proceedings of the Advances in Neural Information Processing Systems, 2017: 6000-6010.

[18] Chen MH, Laina I, Vedaldi A. Training-free layout control with cross-attention guidance [C] // Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision, 2024: 5343-5353.

[19] Blattmann A, Dockhorn T, Kulal S, et al. Stable video diffusion: scaling latent video diffusion models to large datasets [Z/OL]. arXiv Preprint, arXiv: 2311.15127, 2023.

[20] An J, Zhang SY, Yang H, et al. Latent-shift: latent diffusion with temporal shift for efficient text-to-video generation [Z/OL]. arXiv Preprint, arXiv: 2304.08477, 2023.

[21] Blattmann A, Rombach R, Ling H, et al. Align your latents: high-resolution video synthesis with latent diffusion models [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 22563-22575.

[22] Girdhar R, Singh M, Brown A, et al. Emu video: factorizing text-to-video generation by explicit image conditioning [Z/OL]. arXiv Preprint, arXiv: 2311.10709, 2023.

[23] Yu LJ, Cheng Y, Sohn K, et al. MAGVIT: masked generative video Transformer [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 10459-10469.

[24] Gupta A, Yu LJ, Sohn K, et al. Photorealistic video generation with diffusion models [C] // Proceedings of the European Conference on Computer Vision, 2024: 393-411.

[25] Yu LJ, Lezama J, Gundavarapu NB, et al. Language model beats diffusion-tokenizer is key to visual generation [C] // Proceedings of the Twelfth International Conference on Learning Representations, 2024: 1-19.

[26] Polyak A, Zohar A, Brown A, et al. Movie Gen: a cast of media foundation models [Z/OL]. arXiv Preprint, arXiv: 2410.13720, 2024.

[27] Chen HX, Zhang Y, Cun XD, et al. VideoCrafter2: overcoming data limitations for high-quality video diffusion models [C] // Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 7310-7320.

[28] Khachatryan L, Movsisyan A, Tadevosyan V, et al. Text2Video-Zero: text-to-image diffusion models are zero-shot video generators [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 15954-15964.

[29] Wang X, Zhang SW, Yuan HJ, et al. A recipe for scaling up text-to-video generation with text-free videos [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 6572-6582.

[30] Qing ZW, Zhang SW, Wang JY, et al. Hierarchical spatio-temporal decoupling for text-to-video generation [C] // Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 6635-6645.

[31] Menapace W, Siarohin A, Skorokhodov I, et al. Snap video: scaled spatiotemporal Transformers for text-to-video synthesis [C] // Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 7038-7048.

[32] Wu FX, Liu L, Hao FS, et al. Text-to-image synthesis based on object-guided joint-decoding

Transformer [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 18113-18122.

[33] 代婷婷, 范菁, 曲金帅, 等. 基于 Transformer 和对比学习的文本生成图像方法 [J]. 中国科技论文, 2023, 18(7): 793-798.
Dai TT, Fan J, Qu JS, et al. Text generation image method based on Transformer and contrast learning [J]. China Sciencepaper, 2023, 18(7): 793-798.

[34] Skorokhodov I, Menapace W, Siarohin A, et al. Hierarchical patch diffusion models for high-resolution video generation [C] // Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 7569-7579.

[35] Sun MZ, Wang WN, Qin ZH, et al. GLOBER: coherent non-autoregressive video generation via global guided video decoder [C] // Proceedings of the Advances in Neural Information Processing Systems, 2024: 76120-76136.

[36] Li H. Animate anyone: consistent and controllable image-to-video synthesis for character animation [C] // Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 8153-8163.

[37] Shi XY, Huang ZY, Wang FY, et al. Motion-I2V: consistent and controllable image-to-video generation with explicit motion modeling [C] // Proceedings of the ACM SIGGRAPH 2024 Conference Papers, 2024: 1-11.

[38] Ni HM, Egger B, Lohit S, et al. TI2V-Zero: zero-shot image conditioning for text-to-video diffusion models [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 9015-9025.

[39] Chefer H, Zada S, Paiss R, et al. Still-Moving: customized video generation without customized video data [Z/OL]. arXiv Preprint, arXiv: 2407.08674, 2024.

[40] Cheang CL, Chen GZ, Jing Y, et al. GR-2: a generative video-language-action model with Web-scale knowledge for robot manipulation [Z/OL]. arXiv Preprint, arXiv: 2410.06158, 2024.

[41] Heusel M, Ramsauer H, Unterthiner T, et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium [C] // Proceedings of the Advances in Neural Information Processing Systems, 2017: 1-12.

[42] Unterthiner T, Van Steenkiste S, Kurach K, et al. Towards accurate generative models of video: a new metric & challenges [Z/OL]. arXiv Preprint, arXiv: 1812.01717, 2018.

[43] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision [C] // Proceedings of the 38 th International Conference on Machine Learning, 2021: 8748-8763.

[44] Esser P, Chiu J, Atighehchian P, et al. Structure and content-guided video synthesis with diffusion models [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 7346-7356.

[45] Saito M, Saito S, Koyama M, et al. Train sparsely, generate densely: memory-efficient unsupervised training of high-resolution temporal GAN [J]. International Journal of Computer Vision, 2020, 128(10): 2586-2606.

[46] Wang Z, Bovik AC, Sheikh HR, et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[47] Hore A, Ziou D. Image quality metrics: PSNR vs. SSIM [C] // Proceedings of the 2010 20th International Conference on Pattern Recognition, 2010: 2366-2369.

[48] Zhang R, Isola P, Efros AA, et al. The unreasonable effectiveness of deep features as a perceptual metric [C] // Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 586-595.

[49] Ren WM, Yang H, Zhang G, et al. ConsistI2V: enhancing visual consistency for image-to-video generation [J]. Transactions on Machine Learning Research, 2024, 07: 1-28.

[50] Huang ZQ, He YN, Yu JS, et al. VBench: comprehensive benchmark suite for video generative models [C] // Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 21807-21818.

[51] Zheng CX, Vedaldi A. Free3D: consistent novel view synthesis without 3D representation [C] // Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 9720-9731.

[52] Xu DJ, Nie WL, Liu C, et al. CamCo: camera-controllable 3D-consistent image-to-video generation [Z/OL]. arXiv Preprint, arXiv: 2406.02509, 2024.

[53] He H, Xu YH, Guo YW, et al. CameraCtrl: enabling camera control for text-to-video generation [Z/OL]. arXiv Preprint, arXiv: 2404.02101, 2024.

[54] Liu L, Liu QD, Qian SJ, et al. Text-Animator: controllable visual text video generation [Z/OL]. arXiv Preprint, arXiv: 2406.17777, 2024.

[55] Lu Y, Zhu LC, Fan HH, et al. FlowZero: zero-shot text-to-video synthesis with LLM-driven dynamic scene syntax [Z/OL]. arXiv Preprint, arXiv: 2311.15813, 2023.

[56] Long FC, Qiu ZF, Yao T, et al. VideoStudio: generating consistent-content and multiscene videos [C] // Proceedings of the European Conference on Computer Vision, 2024: 468-485.

[57] Wu JZ, Li XT, Zeng YH, et al. MotionBooth: motion-aware customized text-to-video generation [Z/OL]. arXiv Preprint, arXiv: 2406.17758, 2024.

[58] Guo YW, Yang CY, Rao AY, et al. AnimateDiff: animate your personalized text-to-image diffusion models without specific tuning [C] // Proceedings of the Twelfth International Conference on Learning Representations, 2024: 1-19.

[59] Chen XY, Wang YH, Zhang LJ, et al. SEINE: short-to-long video diffusion model for generative transition and prediction [C] // Proceedings of the Twelfth International Conference on Learning Representations, 2023: 1-15.

[60] Yang SY, Hou L, Huang HB, et al. Direct-a-Video: customized video generation with user-directed camera movement and object motion [C] // Proceedings of the ACM SIGGRAPH 2024 Conference Papers, 2024: 1-12.

[61] Wang ZX, Yuan ZY, Wang XT, et al. MotionCtrl: a unified and flexible motion controller for video generation [C] // Proceedings of the ACM SIGGRAPH 2024 Conference Papers, 2024: 1-11.

[62] Ling PY, Bu JZ, Zhang P, et al. MotionClone: training-free motion cloning for controllable video generation [Z/OL]. arXiv Preprint, arXiv: 2406.05338, 2024.

[63] Xie YM, Yao CH, Voleti V, et al. SV4D: dynamic 3D content generation with multi-frame and multi-view consistency [Z/OL]. arXiv Preprint, arXiv: 2407.17470, 2024.

[64] Wu T, Zhang Y, Wang XT, et al. CustomCrafter: customized video generation with preserving motion and concept composition abilities [Z/OL]. arXiv Preprint, arXiv: 2408.13239, 2024.

[65] Hong S, Seo J, Shin H, et al. DirecT2V: large language models are frame-level directors for zero-shot text-to-video generation [C] // Proceedings of the First Workshop on Controllable Video Generation@ ICML24, 2024: 1-16.

[66] Roush A, Zakirov E, Shirokov A, et al. LLM as an Art Director (LaDi): using LLMs to improve Text-to-Media generators [Z/OL]. arXiv preprint arXiv: 2311.03716, 2023.

[67] Xing Z, Dai Q, Weng ZJ, et al. AID: adapting Image2Video diffusion models for instruction-guided video prediction [Z/OL]. arXiv Preprint, arXiv: 2406.06465, 2024.

[68] Gao KF, Shi JX, Zhang HW, et al. ViD-GPT: introducing GPT-style autoregressive generation in video diffusion models [Z/OL]. arXiv Preprint, arXiv: 2406.10981, 2024.

[69] Liu SW, Ren ZZ, Gupta S, et al. PhysGen: rigid-body physics-grounded image-to-video generation [C] // Proceedings of the European Conference on Computer Vision, 2024: 360-378.

[70] Ma Z, Zhou DQ, Yeh CH, et al. Magic-Me: identity-specific video customized diffusion [Z/OL]. arXiv Preprint, arXiv: 2402.09368, 2024.

[71] He XH, Liu QD, Qian SJ, et al. ID-animator: zero-shot identity-preserving human video generation [Z/OL]. arXiv Preprint, arXiv: 2404.15275, 2024.

[72] Gu YC, Zhou YP, Wu BC, et al. VideoSwap: customized video subject swapping with interactive semantic point correspondence [C] //

Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 7621-7630.

[73] Chen H, Wang X, Zhang YP, et al. DisenStudio: customized multi-subject text-to-video generation with disentangled spatial control [Z/OL]. arXiv Preprint, arXiv: 2405.12796, 2024.

[74] Cheng J, Wu FX, Tian YL, et al. RiFeGAN: rich feature generation for text-to-image synthesis from prior knowledge [C] // Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10908-10917.

[75] Huang HZ, Feng YF, Shi C, et al. Free-Bloom: zero-shot text-to-video generator with LLM director and LDM animator [C] // Proceedings of the Advances in Neural Information Processing Systems, 2024: 26135-26158.

[76] Wang AD, Wu B, Chen SL, et al. SOK-Bench: a situated video reasoning benchmark with aligned open-world knowledge [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 13384-13394.

[77] 吴昊文, 王鹏, 李亮亮, 等. 基于语义增强和特征融合的文本生成图像方法 [J/OL]. 计算机工程与应用 (2024-10-17)[2024-12-01]. https://kns.cnki.net/kcms2/article/abstract?v=kz9ikNiCkdrKwwwFaUR9AvFI1w9gaKFGXMS3fkMNsyscvy6p4FWsLT_vKZr9xJdKcrtEE-YHp5byRMw9cbNCqn2Y1nNyZNm38OYdnhNqisI2V1T-xWbqVHJvPR9qQg-Avvd4kKBdXUWV2RVsEhR6uigHyeb6TAx8HKSmR0_HXL549TqZiyISnyWg0SKChaj3&uniplatform=NZKPT&language=CHS.
Wu HW, Wang P, Li LL, et al. Text-to-image generation method based on semantic enhancement and feature fusion [J/OL]. Computer Engineering and Applications (2024-10-17)[2024-12-01]. https://kns.cnki.net/kcms2/article/abstract?v=kz9ikNiCkdrKwwwFaUR9AvFI1w9gaKFGXMS3fkMNsyscvy6p4FWsLT_vKZr9xJdKcrtEE-YHp5byRMw9cbNCqn2Y1nNyZNm38OYdnhNqisI2V1T-xWbqVHJvPR9qQg-Avvd4kKBdXUWV2RVsEhR6uigHyeb6TAx8HKSmR0_HXL549TqZiyISnyWg0SKChaj3&uniplatform=NZKPT&language=CHS.

[78] 李源凡, 张丽红. 基于 CLIP 模型和文本重建的人脸图像生成方法研究 [J]. 测试技术学报, 2024, 38(2): 154-160.
Li YF, Zhang LH. Research on face image generation method based on CLIP model and text reconstruction [J]. Journal of Test and Measurement Technology, 2024, 38(2): 154-160.

[79] 马军, 车进, 贺愉婷, 等. 基于空间注意力及条件增强的文本生成图像方法 [J]. 山东大学学报(工学版) (2024-10-18)[2024-12-01]. https://kns.cnki.net/kcms2/article/abstract?v=kz9ikNiCkdqV31_LzontIXn54JpXZVbKi23HClQBqul7DDrc-c3RnyUHPcLcxxaHrGlaaSZuxUOcCDSfbGQeUcvUABJBcw0auhKsq806gdlYpAumehIP9I1xe3Bg0TmZucKu8rUypHtRE6vdsboctZwcNsOwUGQZziD4Un9bzH-LndQL6bjCxZXxuMzwt_AT&uniplatform=NZKPT&language=CHS.
Ma J, Che J, He YT, et al. Text-to-images synthesis method based on spatial attention and conditional augmentation [J]. Journal of Shandong University(Engineering Science) (2024-10-18)[2024-12-01]. https://kns.cnki.net/kcms2/article/abstract?v=kz9ikNiCkdqV31_LzontIXn54JpXZVbKi23HClQBqul7DDrc-c3RnyUHPcLcxxaHrGlaaSZuxUOcCDSfbGQeUcvUABJBcw0auhKsq806gdlYpAumehIP9I1xe3Bg0TmZucKu8rUypHtRE6vdsboctZwcNsOwUGQZziD4Un9bzH-LndQL6bjCxZXxuMzwt_AT&uniplatform=NZKPT&language=CHS.

[80] 聂开琴, 倪郑威. 基于多文本描述的图像生成方法 [J]. 电信科学, 2024, 40(5): 73-85.
Nie KQ, Ni ZW. Image synthesis method based on multiple text description [J]. Telecommunications Science, 2024, 40(5): 73-85.

[81] Wu FX, Cheng J, Wang XC, et al. Image hallucination from attribute pairs [J]. IEEE Transactions on Cybernetics, 2022, 52(1): 568-581.