自动化学报
ACTA AUTOMATICA SINICA

3D空间先验驱动的相机轨迹可控视频扩散模型

朱泓舟 杨雪 赵敏 李崇轩 朱军

3D Spatial Prior-guided Video Diffusion Models With Controllable Camera Trajectories

ZHU Hong-Zhou, YANG Xue, ZHAO Min, LI Chong-Xuan, ZHU Jun

在线阅读 View online: https://doi.org/10.16383/j.aas.c250124

3D 空间先验驱动的相机轨迹可控视频扩散模型

朱泓舟[1] 杨雪[2] 赵敏[1] 李崇轩[3] 朱军[1]

摘 要 近年来, 视频扩散模型在相机可控的图像到视频生成任务中取得突破性进展. 然而, 现有方法在维持 3D 空间结构一致性方面仍面临显著挑战, 其生成视频普遍存在空间结构模糊化、多视角下物体形态畸变等缺陷, 这些问题严重制约了生成视频的视觉可信度. 为解决这些问题, 提出在视频扩散模型的训练和推理阶段均引入额外的 3D 空间先验信息, 以增强生成视频的空间结构一致性. 具体而言, 在训练阶段, 设计基于视角形变映射的条件嵌入方法 (Warp-Injection), 通过进行逐帧视角形变映射与图像补全构建具备高度空间一致性的参考帧序列, 并将其作为结构先验条件嵌入扩散模型的训练过程. 在推理阶段, 首先提出初始噪声空间几何校正策略 (Warp-Init): 对条件图像加噪进行首帧初始化, 此后通过迭代式视角形变映射构建符合 3D 一致性约束的初始噪声序列. 在此基础上, 进一步在去噪过程中引入基于视角形变先验的能量函数引导策略 (Warp-Guidance), 通过减小生成帧与视角形变映射后的预期目标视频之间的距离来实现对视频 3D 空间一致性的校正. 在标准 RealEstate10K 数据集上的实验结果表明, 相较于当前最优模型, 本文方法在 FVD 指标上取得 18.03 的显著优化, 同时将 3D 结构估计的失败率 (COLMAP error rate) 降低至 5.2%. 可视化分析进一步证明, 本文方法能有效维持生成视频的 3D 空间结构一致性.

关键词 图像到视频生成; 可控生成; 相机轨迹; 扩散模型

引用格式 朱泓舟, 杨雪, 赵敏, 李崇轩, 朱军. 3D 空间先验驱动的相机轨迹可控视频扩散模型. 自动化学报, 2026, **52**(8): 1−12

DOI 10.16383/j.aas.c250124 **CSTR** 32138.14.j.aas.c250124

3D Spatial Prior-guided Video Diffusion Models With Controllable Camera Trajectories

ZHU Hong-Zhou[1] YANG Xue[2] ZHAO Min[1] LI Chong-Xuan[3] ZHU Jun[1]

Abstract In recent years, video diffusion models have achieved breakthrough progress on camera-controllable image-to-video generation. However, existing methods still face significant challenges in maintaining 3D spatial structural consistency: The generated videos commonly exhibit spatial-structure blurring and cross-view shape distortions of objects, which severely constrain their visual credibility. To address these questions, we introduce additional 3D spatial prior information into both the training and inference stages of video diffusion models to enhance the spatial structural consistency of generated videos. Specifically, during the training stage, we design a conditional embedding method based on warping (Warp-Injection): We perform sequential per-frame warping-and-inpainting to construct a reference-frame sequence with high spatial consistency, and embed it as a structural-prior condition into the diffusion model′s training process. During the inference stage, we first propose an initial-noise spatial-geometry correction strategy (Warp-Init): The first frame is initialized by adding noise to the conditional image, and then an initial noise sequence that satisfies 3D-consistency constraints is constructed via iterative warping. On this basis, we further introduce an energy-function guidance strategy (Warp-Guidance) driven by warping priors during denoising, which corrects the video′s 3D spatial consistency by reducing the distance between the generated frames and the expected target video obtained after warping. Experimental results on the standard RealEstate10K dataset show that, compared with the current state-of-the-art model, our method achieves a significant optimization of 18.03 in FVD and reduces the failure rate of 3D structure estimation (COLMAP error rate) to 5.2%. Visualization analyses further demonstrate that our method effectively maintains 3D spatial structural consistency in generated videos.

Keywords image-to-video generation; controllable generation; camera trajectory; diffusion models

Citation Zhu Hong-Zhou, Yang Xue, Zhao Min, Li Chong-Xuan, Zhu Jun. 3D spatial prior-guided video diffusion models with controllable camera trajectories. *Acta Automatica Sinica*, 2026, **52**(8): 1−12

收稿日期 2025-03-25 录用日期 2025-08-04
Manuscript received March 25, 2025; accepted August 4, 2025
本文责任编委 左旺孟
Recommended by Associate Editor ZUO Wang-Meng
1. 清华大学计算机科学与技术系 北京 100084 2. 北京理工大学计算机学院 北京 100081 3. 中国人民大学高瓴人工智能学院 北京 100872
1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084 2. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081 3. Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872

视频生成模型作为当前人工智能领域的重要研究课题之一, 旨在通过对动态场景进行有效且逼真的建模来模拟真实世界的运动过程, 在诸多领域展现出广泛的应用潜力. 与此同时, 为满足用户日益增长的个性化生成需求, 如何基于视频生成基础模型实现可控视频生成逐渐成为研究热点[1−7]. 其一个重要分支是相机轨迹可控的图生视频任务 (camera-controllable image-to-video), 这一技术允许用户以多样化的视角观察物体或场景, 在虚拟现实、电影动画制作与社交媒体内容创作等方面具有重要意义. 该任务旨在以一张静态图像为基础, 结合用户提供的相机轨迹条件, 生成具备特定镜头语言的视频内容[8−13]. 这一任务的一个关键要求是确保生成的视频具备 3D 空间结构一致性.

已有的研究工作 [8−13] 通常着力于设计扩散模型结构, 以增强生成的视频对相机轨迹条件的遵循程度, 并已取得初步成效. 例如, MotionCtrl[8] 设计适配器模块 (adapter) 以便将相机轨迹参数引入模型, 经适配器编码后的特征向量被添加到扩散模型的隐藏特征中; CameraCtrl[9] 设计相机轨迹编码器, 所编码的相机轨迹的特征向量通过时间注意力机制 (temporal attention) 被引入扩散模型中; CamCo[10] 和 CamI2V[11] 则在模型内加入极线约束的注意力模块 (epipolar constraint attention), 通过相机轨迹来对扩散模型的隐藏特征进行几何约束, 以增强模型对相机轨迹变化的感知能力. 然而, 这些方法大多忽视视频生成过程中的 3D 空间结构一致性问题, 导致生成的视频在不同视角间常常存在空间结构模糊化、多视角下物体形态畸变等问题, 严重影响视频的视觉可信度 (见图 1).

为解决这一问题, 本文提出在视频扩散模型的训练和推理阶段均引入额外的 3D 空间先验信息, 以增强生成视频的空间结构一致性. 具体而言, 在模型训练阶段, 设计基于视角形变映射的条件嵌入方法 (Warp-Injection), 通过进行逐帧视角形变映射与图像补全构建具备高度空间一致性的参考帧序列, 并将其作为结构先验条件嵌入扩散模型的训练过程. 为进一步提升空间结构一致性, 本文在扩散

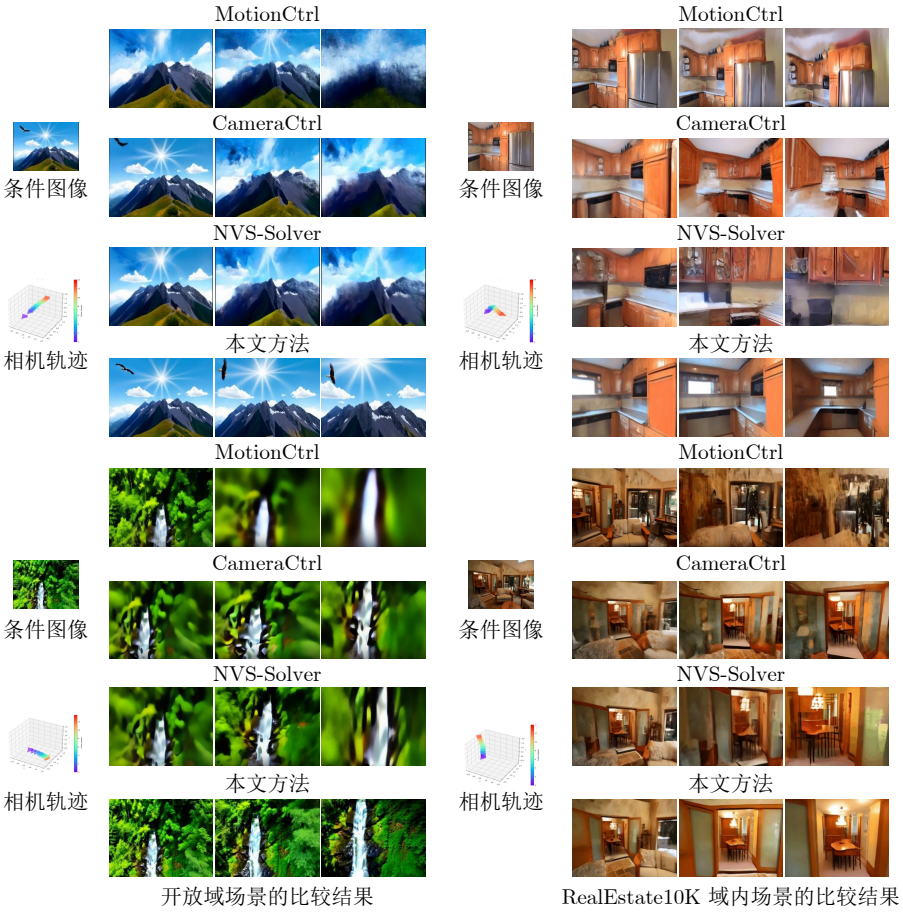

图 1 本文方法与已有的先进方法的定性比较

Fig. 1 Qualitative comparison between our method and existing state-of-the-art methods

模型推理阶段的噪声初始化阶段和去噪阶段, 分别引入初始噪声空间几何校正策略 (Warp-Init) 和基于视角形变先验的能量函数引导策略 (Warp-Guidance). 其中, Warp-Init 在对条件图像加噪进行首帧初始化后, 通过逐帧视角形变映射构建符合 3D 一致性约束的初始噪声序列, 从而赋予初始噪声良好的空间结构; Warp-Guidance 则以生成帧与其经视角形变映射后所得的预期目标视频之间的差异为基础定义能量函数, 鼓励生成帧接近经视角形变映射后的目标帧, 以实现对视频 3D 空间一致性的校正.

为验证所提方法的有效性, 本文在领域内广泛使用的基准数据集 RealEstate10K[14] 上进行全面的实验评估. 实验采用多项定量指标进行评估, 包括 FID (Fréchet inception distance)[15]、FVD (Fréchet video distance)[16]、3D 结构估计的失败率 (COLMAP error rate)[17]、相机平移误差 (translation error) 以及相机旋转误差 (rotation error). 实验结果表明, 相比于已有的先进方法, 本文方法在所有评估指标上均显著优于当前先进模型. 具体而言, 在 FVD 指标上取得 18.03 的优化, 并将 COLMAP error rate 降低至 5.2%, 这充分验证了本文方法在增强视频 3D 空间结构一致性方面的有效性与先进性.

1　相关工作

1.1　用于可控视频生成的扩散模型

扩散模型作为新兴的生成模型, 在图像和视频生成方面取得了显著进展. 这类模型通过前向扩散过程逐渐扰动数据 $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$, 该过程可表示为

$$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0},\, \boldsymbol{I}), \quad t \in [0,\, T] \tag{1}$$

其中, $\boldsymbol{\epsilon}$ 为标准高斯噪声. 这一前向扩散过程可被视为随机微分方程, 其中 α_t 和 σ_t 被称为噪声强度 (noise schedule), 其具体取值能够确保 $\boldsymbol{x}_T$ 的分布接近高斯分布, 即 $p_T(\boldsymbol{x}_T) \sim \mathrm{N}(\boldsymbol{0},\, \boldsymbol{I})$.

鉴于扩散模型在图像生成任务中的卓越表现, 众多研究 [18−25] 探究了其在视频生成中的应用, 其中具有代表性的视频生成模型包括 Stable Video Diffusion (SVD)[26]、Sora[27]、Vidu[28] 等.

视频生成任务的一个重要分支是图像到视频生成 (图生视频)[26, 29−32]. 给定一张图像 I_1, 图像到视频生成的目标是从 I_1 生成一个视频 $X_0 = \{\boldsymbol{x}_0^i\}_{i=1}^N$, 其中保留条件图像 I_1 的视觉元素.

为增强视频生成的可控性, 诸多研究 [1, 3–5, 21] 为扩散模型引入可控机制. 可控生成任务要求训练扩散模型近似一个条件数据分布 $p(X_0|c)$, 其中 c 为控制条件. 模型通过噪声预测网络 $\boldsymbol{\epsilon}_\theta(X_t,\, I_1,\, c,\, t)$[33] 进行参数化, 通过最小化以下训练目标进行训练:

$$\mathrm{E}_{X_0,\, I_1,\, c,\, \boldsymbol{\epsilon},\, t}\left[\|\boldsymbol{\epsilon}_\theta(X_t,\, I_1,\, c,\, t) - \boldsymbol{\epsilon}\|^2\right] \tag{2}$$

其中, 带噪输入 X_t 由式 (1) 产生.

在推理阶段, 可通过从 $X_T \sim p_T(X_T)$ 开始逆转前向过程生成样本. 在条件可控的生成任务中, 为提升所生成视频的视觉质量与对控制条件的遵循程度, 无分类器引导策略 (classifier-free guidance)[34] 被广泛使用. 这一策略中, 推理阶段所使用的去噪项为:

$$\tilde{\boldsymbol{\epsilon}}_\theta(X_t,\, I_1,\, c,\, t) = (1+\gamma)\boldsymbol{\epsilon}_\theta(X_t,\, I_1,\, c,\, t) - \gamma\boldsymbol{\epsilon}_\theta(X_t,\, t) \tag{3}$$

其中, $\gamma > 0$ 控制条件引导的强度大小.

为进一步提升可控视频生成的可控性, 部分研究提出了推理过程中的能量引导 (energy guidance) 策略[35–36], 以在推理过程中引导模型生成更加符合控制条件的视频内容. 这一策略须设计与控制条件 c 密切相关的能量函数 $\mathcal{E}(X_t,\, c,\, t)$, 并将其梯度 $\nabla_{X_t}\mathcal{E}(X_t,\, c,\, t)$ 加入去噪项中, 所得去噪项可表示为:

$$\begin{aligned}\hat{\boldsymbol{\epsilon}}_\theta(X_t,\, I_1,\, c,\, t) = (1+\gamma)\boldsymbol{\epsilon}_\theta(X_t,\, I_1,\, c,\, t) - \\ \gamma\boldsymbol{\epsilon}_\theta(X_t,\, t) + \lambda\nabla_{X_t}\mathcal{E}(X_t,\, c,\, t)\end{aligned} \tag{4}$$

其中, $\lambda > 0$ 控制能量引导的强度大小.

1.2　相机轨迹可控的图像到视频生成

可控视频生成任务的一个重要分支是相机轨迹可控的图生视频任务. 这一任务旨在根据一系列目标相机轨迹 $c = \{c_i\}_{i=1}^N$ 和条件图像 I_1 生成视频 $X_0 = \{\boldsymbol{x}_0^i\}_{i=1}^N$.

早期工作 [21, 37] 实现了对特定相机轨迹的控制. 例如, AnimateDiff[21] 针对 8 种特定的相机轨迹分别训练 LoRA[38] 模块; VideoComposer[37] 则引入二维的运动向量, 使模型能够识别简单的相机运动.

为实现更复杂的相机轨迹控制, MotionCtrl[8] 提出将相机轨迹条件参数化为相机参数矩阵. 具体而言, 相机轨迹被参数化为 $c_i = (\boldsymbol{K}_i,\, \boldsymbol{R}_i,\, \boldsymbol{t}_i)$, 其中 $\boldsymbol{K}_i \in \mathbf{R}^3$、$\boldsymbol{R}_i \in \mathbf{R}^3$ 和 $\boldsymbol{t}_i \in \mathbf{R}^3$ 分别表示相机内参、相机旋转外参和相机平移外参. 为进一步提高模型对相机轨迹条件的理解能力, CameraCtrl[9] 提出使用普吕克嵌入编码 (Plücker embedding) 表示相机轨迹. 该编码基于原始相机参数, 通过下式计算而得:

$$c = \{\ddot{\boldsymbol{P}}_i\}_{i=1}^N, \quad \ddot{\boldsymbol{P}}_i \in \mathbf{R}^{6\times H\times W} \tag{5}$$

其中, H, W 为图像的高度和宽度. $\ddot{\boldsymbol{P}}_i$ 的 $(h,\ w)$ 分量为

$$\ddot{\boldsymbol{P}}_i(h,\ w) = (\boldsymbol{t}_i \times \hat{d}_i(h,\ w),\ \hat{d}_i(h,\ w)) \in \mathbf{R}^6 \tag{6}$$

$$\hat{d}_i(h,\ w) = \frac{d_i(h,\ w)}{\|d_i(h,\ w)\|} \in \mathbf{R}^3 \tag{7}$$

$$d_i(h,\ w) = \boldsymbol{R}_i \boldsymbol{K}_i^{-1} \begin{bmatrix} h \\ w \\ 1 \end{bmatrix} + \boldsymbol{t}_i \in \mathbf{R}^3 \tag{8}$$

这一表示蕴含了相机轨迹的几何特征, 从而增强了模型对相机轨迹条件的感知能力. 为进一步提高生成视频的质量, CamCo[10] 和 CamI2V[11] 在模型内加入极线约束的注意力模块, 通过相机轨迹来对扩散模型的隐藏特征进行几何约束, 以增强模型对相机轨迹变化的感知能力. 此外, 一些工作还提出了无需训练的相机轨迹控制方法[12−13].

2 本文方法

尽管相机轨迹可控的图生视频任务已取得显著进展, 但如图 1 所示, 现有方法生成的视频在 3D 空间结构一致性方面仍有不足, 具体表现为生成的视频存在空间结构模糊化、多视角下物体形态畸变等缺陷, 这严重制约了生成视频的视觉可信度. 为解决这一问题, 本文提出向视频扩散模型中引入额外的空间先验, 从而提升视频的 3D 空间结构一致性.

考虑到基于深度的视角形变映射方法 (depth-based image warping)[39] 能够基于 3D 图形学原理, 构建具有高度空间一致性的图像序列, 本文利用这一方法来为模型提供 3D 空间先验. 该方法定义了视角形变映射函数 $\mathcal{W}$, 能够基于给定图像 I 的深度图 (depth map) $\boldsymbol{D}$ 和相机视角, 将原始图像中特定位置的像素映射到目标视角, 从而构建目标相机视角对应的新视角图像. 具体而言, 设源图像 I 的相机视角为 $c_s = (\boldsymbol{K}_s,\ \boldsymbol{R}_s,\ \boldsymbol{t}_s)$, 目标相机视角为 $c_t = (\boldsymbol{K}_t,\ \boldsymbol{R}_t,\ \boldsymbol{t}_t)$, 则 $\mathcal{W}$ 可将 I 从源相机视角 c_s 映射到目标视角 c_t, 其具体映射方式为:

$$\mathcal{W}(I,\ c_t,\ c_s,\ \boldsymbol{D})(x_t,\ y_t) = I(x_s,\ y_s) \tag{9}$$

在这一映射过程中, 源图像 I 中坐标 $(x_s,\ y_s)$ 处的像素值被映射到目标视角图像的 $(x_t,\ y_t)$ 坐标处. 其中, 对于目标视角图像中 $(x_t,\ y_t)$ 处的像素, 其源视角对应像素位置 $(x_s,\ y_s)$ 可根据下式获得:

$$\begin{bmatrix} \hat{x}_s \\ \hat{y}_s \\ w \end{bmatrix} = \boldsymbol{K}_s \left(\boldsymbol{R}_s \boldsymbol{R}_t^{-1} \left(\boldsymbol{K}_t^{-1} \boldsymbol{D}_{st}(x_t,\ y_t) \begin{bmatrix} x_t \\ y_t \\ 1 \end{bmatrix} - \boldsymbol{t}_t \right) + \boldsymbol{t}_s \right) \tag{10}$$

$$(x_s,\ y_s) = \left(\frac{\hat{x}_s}{w},\ \frac{\hat{y}_s}{w} \right) \tag{11}$$

其中, $\boldsymbol{D}_{st}$ 根据 $\boldsymbol{D}$ 通过前向投影求出. 这一映射过程在图 2 的左上角有所说明.

由于式 (9) 基于基本的 3D 图形学原理, 因此映射所得图像与原始图像保持了高度的 3D 空间结构一致性, 这提供了一个理想的空间先验. 由此, 可利用其来增强生成视频中的 3D 空间结构一致性. 在下文中, 将描述如何在训练 (第 2.1 节) 和推理 (第 2.2 节) 阶段将此空间先验引入相机轨迹可控的图像到视频的扩散模型中.

2.1 基于视角形变映射的条件嵌入方法

为在训练阶段有效地引入 3D 空间先验, 本文提出一种基于视角形变映射的条件嵌入方法 (Warp-Injection), 即, 通过使用视角形变函数 (warping function) 来估计目标视角下的参考帧, 并将之作为条件嵌入扩散模型, 以强化生成视频的 3D 空间结构一致性. 为实现这一目标, 一种简单的方法是从条件图像一次性估计所有目标帧. 但如图 3 所示, 这一策略会导致所估计的目标帧存在较大的无效区域 (图 3 中第一行的黑色区域), 这些无效区域削弱了空间先验的有效性, 从而限制了对模型性能的提升.

为解决这一问题, 本文使用一个预训练的图像补全模型 (inpainting model)[40] 来填充这些无效区域. 然而, 如图 3 第二行所示, 独立地补全无效区域会导致帧间的补全区域不具备空间一致性, 从而无法为模型提供有效先验. 因此, 本文采用一种时序迭代的方法来估计目标帧, 即顺序逐帧进行视角形变映射与图像补全. 具体而言, 如图 2 的上半部分所示, 从条件图像开始, 每一帧均由前一帧经视角形变映射得到, 继而补全视角形变映射所得图像中的无效区域. 如图 3 第三行所示, 这一方法可确保补全区域在帧间保持一致, 从而维持较高的 3D 空间结构一致性. 形式化地, 给定条件图像 I_1 和目标相机轨迹 $\{c_i\}_{i=1}^N$, 每一帧 I_{i+1} 均可通过下式估计得到:

$$I_{i+1} = \mathcal{P}(\mathcal{W}(I_i,\ c_{i+1},\ c_i,\ \mathcal{D}(I_i))) \tag{12}$$

其中, $\mathcal{D}(\cdot)$ 和 $\mathcal{P}(\cdot)$ 分别代表单目深度估计模型和图像补全模型; $\mathcal{W}(\cdot,\ \cdot,\ \cdot,\ \cdot)$ 是定义在式 (9) 中的视角形变映射函数. 尽管这一方法已能获得具备较高空间一致性的估计帧, 但由于在顺序迭代过程中, 每帧的计算过程均依赖之前帧, 因此可能存在误差累积问题. 具体而言, 若所估计的某帧视觉质量较差, 则其后续帧将继承这一低质量内容, 从而导致后续

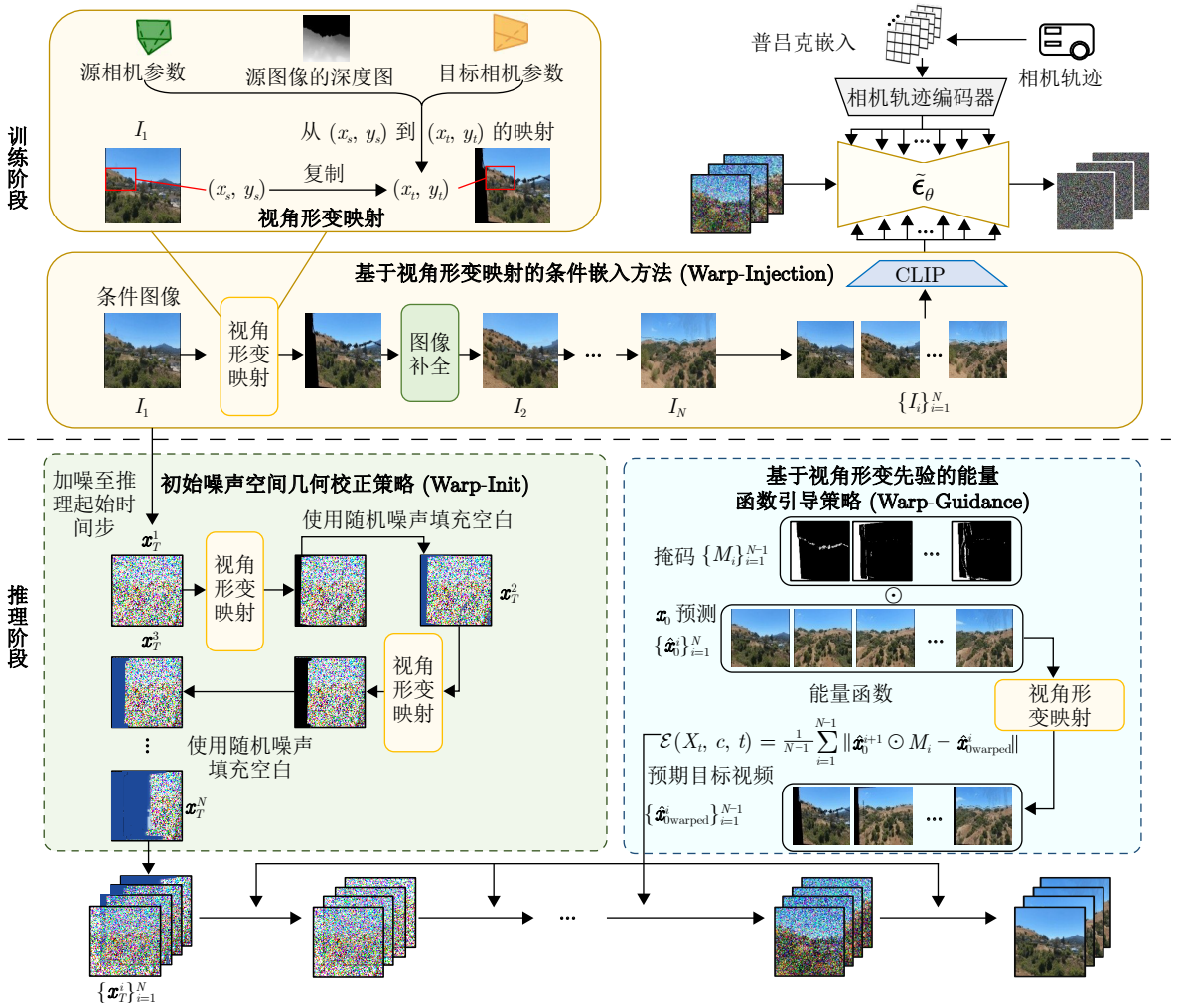

图 2　本文方法结构图

Fig. 2　Structural diagram of our method

估计帧的质量大幅下降. 为解决这一问题, 本文在完成每一帧的估计后, 将其通过视角形变映射函数反向映射到条件图像视角, 计算其有效区域与条件图像的相似程度, 若差别过大, 则改用条件图像直接视角形变映射到这一帧的视角. 形式化地, 本方法先通过以下方式将 I_{i+1} 映射到条件图像视角:

$$I'_{i+1} = \mathcal{W}(I_{i+1},\ c_1,\ c_{i+1},\ \mathcal{D}(I_{i+1})) \tag{13}$$

其中, 形变映射的有效区域掩码为 M. 在此基础上, 计算有效区域中 I'_{i+1} 与条件图像 I_1 的相对差异:

$$\epsilon_r = \frac{||(I_1 - I'_{i+1}) \odot M||}{||I_1 \odot M||} \tag{14}$$

其中, 符号 $\odot$ 表示逐元素乘法. 若相对差异 $\epsilon_r \leq 50\%$, 则仍采用原 I_{i+1} 作为视角 c_{i+1} 的估计帧, 否则采用条件图像直接映射到目标视角并加以补全:

$$I_{i+1} = \mathcal{P}(\mathcal{W}(I_1,\ c_{i+1},\ c_1,\ \mathcal{D}(I_1))) \tag{15}$$

通过采用这一方法, 可在保持帧间 3D 空间结构一致性的同时, 避免出现质量过差的估计帧, 并缓解顺序迭代的误差累积问题. 本文将这一方法得到的估计帧称为参考帧.

然而, 上述参考帧虽然具备良好的空间结构一致性, 但在视觉细节上仍然存在不足, 如存在缺乏高质量细节、视觉效果过饱和等问题, 因此无法直接作为最终的视频结果. 此外, 由于直接将参考帧作为控制条件引入扩散模型会导致上述低质量细节被同时引入, 因此参考帧无法直接作为控制条件. 考虑到 CLIP 模型[41] 能够提取图像的抽象特征且会忽略细粒度的细节, 本文采用 CLIP 模型[41] 提取参考帧的空间结构特征, 并通过交叉注意力机制引入到扩散模型. 由于这一策略在捕捉高层次语义信息的同时能够忽略细粒度的细节, 因此模型能够有效利用参考帧的 3D 空间结构一致性特征, 且不会受到参考帧低质量细节的影响. 在 CLIP 模型完成对

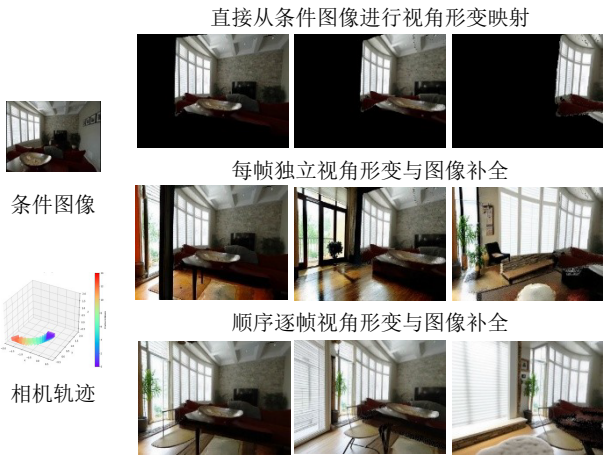

图 3 顺序逐帧视角形变映射与图像补全策略的有效性

Fig. 3 Effectiveness of the sequential per-frame warping-and-inpainting strategy

参考帧的编码后, 所得特征向量将通过交叉注意力机制被引入扩散模型中. 具体而言, 将第 i 个参考帧经 CLIP 模型编码后的特征向量表示为 e_i, 待生成视频的第 i 帧在扩散模型中对应的隐藏特征表示为 v_i, 则扩散模型交叉注意力机制 (cross attention, CA) 的计算过程可描述为:

$$CA(v_i,\ e_i) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}V\right) \tag{16}$$

$$Q = v_i W^Q, \quad K = e_i W^K, \quad V = e_i W^V \tag{17}$$

其中, W^Q, W^K, W^V 为注意力机制中模型的可学习权重; d 为隐藏特征的特征维度.

本文采用 SVD[26] 作为基础模型. 参考 CameraCtrl[9], 本文将相机轨迹编码为普吕克嵌入, 并采用相机轨迹编码器将之嵌入模型, 以提高模型对相机轨迹的遵循程度. 具体而言, 本文采用 CameraCtrl[9] 中的相机条件注入方法, 通过相机姿态编码器将普吕克嵌入编码为特征向量, 将其与扩散模型每层隐藏特征相加, 并通过线性层和时间注意力层使二者更好地融合. 所得隐藏特征向量将通过上文方式与参考帧的特征向量进行交叉注意力计算, 从而使模型集成相机条件与参考帧信息, 生成兼具精确相机轨迹与优质空间结构一致性的视频内容.

上述训练策略综合了视角形变映射的显式相机位姿信息与普吕克嵌入的隐式相机位姿信息, 能够充分结合二者的优势. 具体而言, 视角形变映射能够提供充分的空间结构先验, 但不具备定量的相机位姿信息, 难以精确控制生成视频的相机轨迹; 普吕克嵌入虽能精确反映相机位姿条件, 但缺乏空间结构信息, 不利于生成的视频保持空间一致性. 将二者同时作为条件, 可使生成的视频在精准遵循相机轨迹的同时具备高度空间结构一致性. 尽管这两种条件对同一位姿进行不同处理, 可能存在一定差异, 但在训练过程中, 模型可自适应地优化二者的输入模块, 从而滤除这种差异, 实现两种条件信息的有效结合.

2.2 适用于模型推理阶段的 3D 空间先验约束方法

为进一步提升视频的空间结构一致性, 本文在扩散模型推理阶段的噪声初始化阶段和去噪阶段, 分别引入初始噪声空间几何校正策略 (Warp-Init) 和基于视角形变先验的能量函数引导策略 (Warp-Guidance), 推理算法流程如算法 1 所示. 下面将对这两种策略分别加以介绍.

算法 1. 结合 Warp-Init 和 Warp-Guidance 的相机轨迹可控的图生视频扩散模型采样算法

输入. 起始时间步 T, 噪声预测模型 $\boldsymbol{\epsilon}_\theta(\cdot,\ \cdot,\ \cdot,\ \cdot)$, 去噪采样器 $Sampler(\cdot,\ \cdot,\ \cdot)$, 条件图像 I_1, 条件相机轨迹 $\{c_i\}_{i=1}^N$, 参考帧深度图 $\{\hat{D}_i\}_{i=1}^N$, 视角形变映射函数 $\mathcal{W}(\cdot,\ \cdot,\ \cdot,\ \cdot)$, 式 (20) 中定义的能量函数 $\mathcal{E}(\cdot,\ \cdot,\ \cdot)$.

1) 对条件图像 I_1 加噪至起始时间步 T 以获得第一帧对应的初始化噪声 $\boldsymbol{x}_T^1$;
2) **for** $i = 1 : N-1$ **do**
3) 根据上一帧的初始化噪声计算当前帧的初始化噪声 $\boldsymbol{x}_T^{i+1} = \mathcal{W}(\boldsymbol{x}_T^i,\ c_{i+1},\ c_i,\ \hat{D}_i)$;
4) 用随机噪声填充 $\boldsymbol{x}_T^{i+1}$ 的无效区域; # Warp-Init
5) **end for**
6) 设置初始化噪声 $X_T = \{\boldsymbol{x}_T^i\}_{i=1}^N$;
7) **for** $t = T : 1$ **do**
8) 根据式 (20) 计算能量函数 $\mathcal{E}(X_t,\ c,\ t)$;
9) 根据式 (4) 用噪声预测网络 $\boldsymbol{\epsilon}_\theta(X_t,\ I_1,\ c,\ t)$ 和能量函数 $\mathcal{E}(X_t,\ c,\ t)$ 计算去噪项 $\hat{\boldsymbol{\epsilon}}_\theta(X_t,\ I_1,\ c,\ t)$;
10) 执行去噪过程, 计算 $X_{t-1} = Sampler(X_t,\ \hat{\boldsymbol{\epsilon}}_\theta(X_t,\ I_1,\ c,\ t),\ t)$; # Warp-Guidance
11) **end for**
12) **Return** X_0

2.2.1 初始噪声空间几何校正策略

已有研究 [32, 42–45] 指出, 扩散模型在生成阶段的初始噪声会显著影响最终生成结果. 受此启发, 本文提出利用视角形变映射为初始噪声赋予空间结构信息, 从而在生成过程中引入 3D 空间先验. 受 CamTrol[12] 启发, 一种简单的噪声初始化方法是借助类似于 SDEdit[46] 的思路, 直接在第 2.1 节所获参考帧基础上添加随机噪声, 并以其作为初始噪声. 然而, 如图 4 第二行所示, 这一策略使参考帧中的

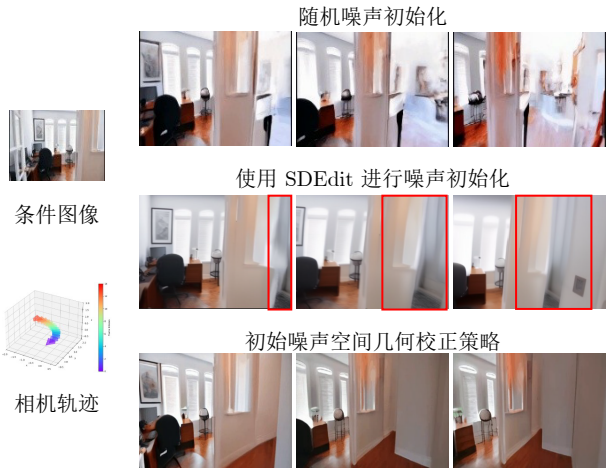

图 4　初始噪声空间几何校正策略 (Warp-Init) 的效果

Fig. 4　Effect of the initial-noise spatial-geometry correction strategy (Warp-Init)

低质量细节信息被保留到最终生成的视频中, 从而降低了视频的视觉质量.

为在去噪过程中引入空间先验的同时避免引入低质量细节, 本节提出一种初始噪声空间几何校正策略 (Warp-Init), 以对初始噪声进行空间几何校正. 具体而言, 这一策略不再直接对全部参考帧添加噪声, 而是通过在初始噪声序列之间执行逐帧视角形变映射, 以间接引入空间结构先验, 从而既能引入所需的 3D 空间结构一致性特征, 又能有效避免引入参考帧中存在的低质量细节信息.

具体而言, 为充分利用条件图像的先验信息, 本方法首先将条件图像加噪到推理起始的时间步, 并以其作为视频第一帧的初始噪声 $\boldsymbol{x}_T^1$. 随后, 利用给定的目标相机轨迹 $\{c_i\}_{i=1}^N$ 以及参考帧对应的深度图 $\{\hat{D}_i\}_{i=1}^N$, 将上一帧的噪声逐帧进行视角形变映射, 以得到下一帧对应视角下的初始化噪声, 从而迭代式地生成噪声序列. 这一过程的形式化表示如下:

$$\boldsymbol{x}_T^{i+1} = \mathcal{W}(\boldsymbol{x}_T^i,\ c_{i+1},\ c_i,\ \hat{D}_i) \tag{18}$$

其中, $\boldsymbol{x}_T^{i+1}$ 表示第 $i+1$ 帧对应的初始噪声. 对于视角形变映射过程中产生的无效区域, 本文方法使用标准高斯随机噪声对其进行填充, 最终获得完整的初始噪声序列 $X_T = \{\boldsymbol{x}_T^i\}_{i=1}^N$, 并将其用于扩散模型的去噪过程. 如图 4 第三行所示, 这一策略能够显著提升生成视频的 3D 空间结构一致性. 同时, 由于这一方法只需调用视角形变映射函数及上文 Warp-Injection 中所得的深度图, 无需调用基于深度学习的模型, 因此不会引入过多推理时延, 不会严重影响生成效率. 具体而言, 以随机高斯噪声为初始化噪声的基础方法, 通过 25 步采样, 在单张 A800 GPU 上生成第 3.1 节所设置分辨率下的视频, 推理用时为 118 s, 而本方法在相同设置下, 推理用时为 121 s, 仅增加 3 s 时延.

2.2.2　基于视角形变先验的能量函数引导策略

在通过 Warp-Init 对初始噪声进行空间几何校正后, 本节进一步在扩散模型的去噪过程中引入空间先验信息. 考虑到能量引导策略 (energy guidance) 能够有效引导模型生成高度遵循控制条件的视频[34−36, 47], 本文提出基于视角形变先验的能量函数引导策略 (Warp-Guidance), 通过减小生成帧与视角形变映射后的预期目标视频之间的距离来实现对视频 3D 空间一致性的校正.

具体而言, 在扩散模型去噪阶段的每个时间步 t, 本文方法首先使用扩散模型进行单步的 $\boldsymbol{x}_0$ 预测 ($\boldsymbol{x}_0$-prediction), 获得当前时刻的视频预测结果 $\{\hat{\boldsymbol{x}}_0^i\}_{i=1}^N$. 随后, 本文方法利用第 2.1 节中估计所得参考帧的深度图 $\{\hat{D}_i\}_{i=1}^N$, 对每一预测帧 $\hat{\boldsymbol{x}}_0^i$ 进行视角形变映射, 以得到下一帧对应相机视角下的图像 $\hat{\boldsymbol{x}}_{0\text{warped}}^i$, 其具体计算过程为

$$\hat{\boldsymbol{x}}_{0\text{warped}}^i = \mathcal{W}(\hat{\boldsymbol{x}}_0^i,\ c_{i+1},\ c_i,\ \hat{D}_i) \tag{19}$$

经视角形变映射所得的估计帧与 $\hat{\boldsymbol{x}}_0^i$ 之间具备高度的空间一致性, 因此可被视为预期的目标视频帧. 本文方法通过减少预测视频与预期目标视频的差异, 来引导模型生成具有高度空间一致性的视频. 形式化地, 本方法计算上述估计帧 $\hat{\boldsymbol{x}}_{0\text{warped}}^i$ 与下一预测帧 $\hat{\boldsymbol{x}}_0^{i+1}$ 之间的 $\mathcal{L}_2$ 距离, 并以此定义能量函数. 考虑到视角形变映射过程中存在无效区域, 本方法引入掩码机制, 以避免无效区域对能量计算的影响. 具体的能量函数 $\mathcal{E}(X_t,\ c,\ t)$ 定义如下:

$$\mathcal{E}(X_t,\ c,\ t) = \frac{1}{N-1}\sum_{i=1}^{N-1}\|\hat{\boldsymbol{x}}_0^{i+1}\odot M_i - \hat{\boldsymbol{x}}_{0\text{warped}}^i\| \tag{20}$$

其中, M_i 是用于标记视角形变映射后的无效区域的二值掩码. 本文方法将该能量函数的梯度代入式 (4) 中, 指导扩散模型的去噪过程. 如图 5 第三行所示, 这一能量引导方法显著提高了视频生成的 3D 空间结构一致性.

此外, 本文还探索了一种更简单的能量函数设计, 类似于 NVS-Solver[13], 只对条件图像 I_1 进行视角形变映射来定义能量函数:

$$\mathcal{E}(X_t,\ c,\ t) = \frac{1}{N}\sum_{i=1}^{N}\|\hat{\boldsymbol{x}}_0^i\odot M_i - \mathcal{W}(I_1,\ c_i,\ c_1,\ \hat{D}_1)\| \tag{21}$$

然而, 如图 5 第二行所示, 这种方法的效果不

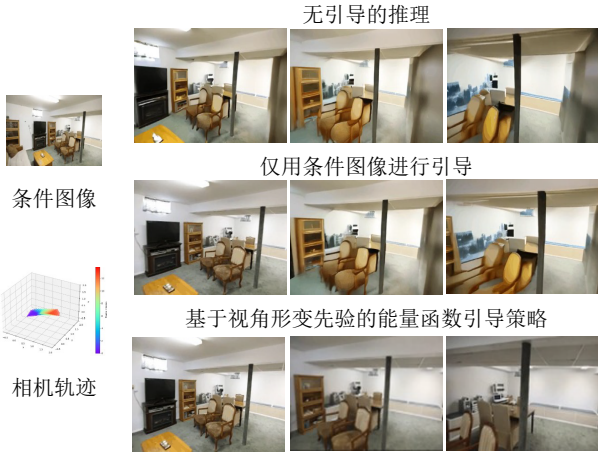

图 5 基于视角形变先验的能量函数引导策略 (Warp-Guidance) 的效果

Fig. 5 Effect of the warping-prior-based energy-function guidance strategy (Warp-Guidance)

尽理想. 这一现象的可能原因是当目标视角与条件图像视角相差较大时, 直接对条件图像进行视角形变映射所产生的无效区域过大, 导致引导作用受到削弱.

3 实验与分析

3.1 实验设置

本文参照 CameraCtrl[9], 在 RealEstate10K[14] 数据集上训练和评估本文所提出的模型, 该数据集由大约 70 000 个带有相应相机轨迹的视频片段组成. 在训练过程中, 本文将视频调整为 320×576 的分辨率, 同时保持原始纵横比, 然后进行中心裁剪以符合本文的训练设置.

本文使用 Depth Anything[48] 进行深度估计, 并应用 Flux-fill[49] 进行图像补全. 本文采用预训练的 CameraCtrl[9] 模型来初始化权重, 以 48 的批大小 (batch size) 训练 10 000 步. 在训练期间, 本文使用 AdamW 优化器: 学习率为 3×10^{-5}; 权重衰减系数为 1×10^{-2}; $\epsilon = 1 \times 10^{-8}$; $\beta_1 = 0.9$; $\beta_2 = 0.999$. 在推理过程中, 对于 Warp-Init, 使用核大小为 8×8、步长为 8 的平均池化来对齐估计深度图与隐空间的维度. 对于 Warp-Guidance, 将式 (4) 中的 γ 和 λ 均设置为 7. 值得注意的是, 在使用 Depth Anything 模型进行深度估计时, 难以精确保证每个点的深度准确性. 然而, 在绝大多数场景中, Depth Anything 已经能够实现较为准确的估计. 本文提出的方法框架具有良好的灵活性, 未来一旦有更先进、更鲁棒的深度估计模型出现, 本文方法可无缝集成, 从而进一步提升性能.

3.2 评估指标

参照文献 [10], 本文采用 COLMAP[17] 来评估生成视频的 3D 空间结构一致性和相机轨迹准确性, 并通过 FID[15] 和 FVD[16] 来评估视频的整体视觉质量. 这些指标的具体说明如下.

COLMAP error rate. COLMAP[17] 是一个从运动估计结构的模型, 能够从视频序列中重建 3D 结构并估计相应的相机轨迹. 对于空间结构退化或 3D 空间结构一致性过差的视频, COLMAP 无法执行三维重建. 因此, 本文采用 COLMAP error rate 作为评估生成视频 3D 空间结构一致性的指标. 考虑到 COLMAP 流程中引入的随机性, 本节对每个视频进行五次尝试, 并且只有当至少一次尝试成功匹配所有 14 帧时, 才认为该视频的 COLMAP 分析是成功的.

相机轨迹准确性. 为评估生成视频的相机轨迹准确性, 本文将 COLMAP 估计所得的相机外参 $\{\boldsymbol{R}^i_{pred}, \boldsymbol{t}^i_{pred}\}_{i=1}^N$ 与真实相机参数 $\{\boldsymbol{R}^i_{gt}, \boldsymbol{t}^i_{gt}\}_{i=1}^N$ 进行比较, 分别计算 Transition error 和 Rotation error 如下:

$$\text{Transition error} = \sum_{i=1}^{N} \|\boldsymbol{t}^i_{pred} - \boldsymbol{t}^i_{gt}\| \tag{22}$$

$$\text{Rotation error} = \sum_{i=1}^{N} \arccos\left(\frac{\text{tr}((\boldsymbol{R}^i_{pred})^{-1}\boldsymbol{R}^i_{gt}) - 1}{2}\right) \tag{23}$$

FID 和 FVD. 为评估视频的整体质量, 本文使用 FID[15] 和 FVD[16] 来测量生成视频与真实视频之间的差异. 在计算 FVD 时, 采用文献 [11] 的 FVD 计算程序; 在计算 FID 时, 将视频的每一帧视为单独的图像.

3.3 性能比较

本节将本文方法与已有的先进方法进行比较, 包括 MotionCtrl[8]、CameraCtrl[9] 和 NVS-Solver[13]. 参考 CameraCtrl[9], 本文从 RealEstate10K[14] 测试数据集中随机抽取 1 000 个视频片段, 并根据第 3.2 节中定义的指标进行评估, 所有结果均为样本平均值.

表 1 和图 1 提供了定量和定性的比较, 其中也展示了开放域情况下的定性比较结果. 如表 1 所示, 本文方法相比所有基线方法均有较高优势. 相较于最优基线方法 (CameraCtrl), 本文方法在表征视频整体质量的指标 FVD 上取得了 18.03 的显著优化, 并取得了低至 5.2% 的 COLMAP error rate. 这表明

表 1　在 RealEstate10K 上的定量比较 (最佳结果以粗体显示)

Table 1　Quantitative comparison on RealEstate10K (the best results are shown in bold)

方法	FID ↓	FVD ↓	COLMAP error rate ↓	Transition error ↓	Rotation error ↓
MotionCtrl[8]	11.22	91.93	22.3%	5.7274	2.5416
CameraCtrl[9]	11.09	70.66	19.8%	5.2519	2.1190
NVS-Solver[13]	12.01	79.82	20.9%	5.1824	2.3491
本文方法	**9.61**	**52.63**	**5.2%**	**3.6372**	**1.9633**

本文方法能够生成兼具优质视觉质量与高度空间一致性的视频. 此外, 在反映相机轨迹准确性的指标 (相机平移误差和旋转误差) 上, 本文方法也取得了最佳结果, 展现出了对相机条件的高度遵循能力 (camera pose alignment). 图 1 中的可视化结果也表明, 在基线方法表现出结构模糊、物体畸变的场景中, 本文方法能够生成具备优良空间一致性的视频. 这一结果与上文所述的定量结果相符合, 充分表明本文方法能够稳定提升相机轨迹可控图生视频任务的空间一致性.

除以上与相机姿态可控的视频生成方法比较外, 本节还将本文方法与 3D 新视角场景生成方法进行比较. 当前先进的 3D 新视角场景生成方法[50–52], 如 ViewCrafter[51] 和 GEN3C[52], 在 3D 场景生成过程中也采用了视频生成模型. 此处将本文方法与 ViewCrafter[51] 等 3D 场景生成方法进行定性对比. 尽管这些方法也采用了视频扩散模型作为关键组成部分, 但其目标在于生成具备高度 3D 一致性的 3D 空间场景, 而非可控视频生成. 因此, 相比本文方法, 这些方法在视频物体的动态性方面有所欠缺, 难以被应用于可控视频生成任务中. 具体而言, 以 FlexWorld[50] 和 ViewCrafter[51] 为代表的 3D 新视角场景生成方法, 大多维护了一个全局立体模型 (dense stereo model), 如点云模型, 并通过视频模型生成的视频对其进行渐进式补全, 视频模型每次迭代仅负责生成一段视角内的静态内容. 如图 6 所示, 由于这一“渐进式补全全局立体模型”的方法是基于“3D 场景中物体保持静止”的基本假设, 因此所渲染出的场景中物体大多保持静止, 这与可控视频生成任务对物体动态性的要求相悖, 不适用于这一任务. 相比之下, 如图 6 所示, 本文方法虽然引入了空间结构先验信息, 但并非基于全局立体模型补全的方式, 因此未对物体动态性造成严重损害.

3.4　消融实验

为评估本文方法的有效性, 本节对 Warp-Injection、Warp-Init、Warp-Guidance 进行消融实验. 本节在与第 3.1 节中相同的设置下训练基线模型, 并评估上述策略的有效性. 如表 2 所示 (表中“Inj.”、“Init.”、“Guid.”分别代表 Warp-Injection、Warp-Init、Warp-Guidance), 相对于基线方法, 使用 Warp-Injection 向扩散模型中引入空间先验信息, 能够显著优化 FVD 等指标, 并将 COLMAP error rate 降至 7.6%. 这表明 Warp-Injection 能够提升生成视频的视觉质量和空间一致性. 在此基础上, 在模型推理阶段分别应用 Warp-Init、Warp-Guidance, 能够进一步提升视频的空间一致性, 其 COLMAP error rate 分别降低至 7.1% 和 6.7%. 综

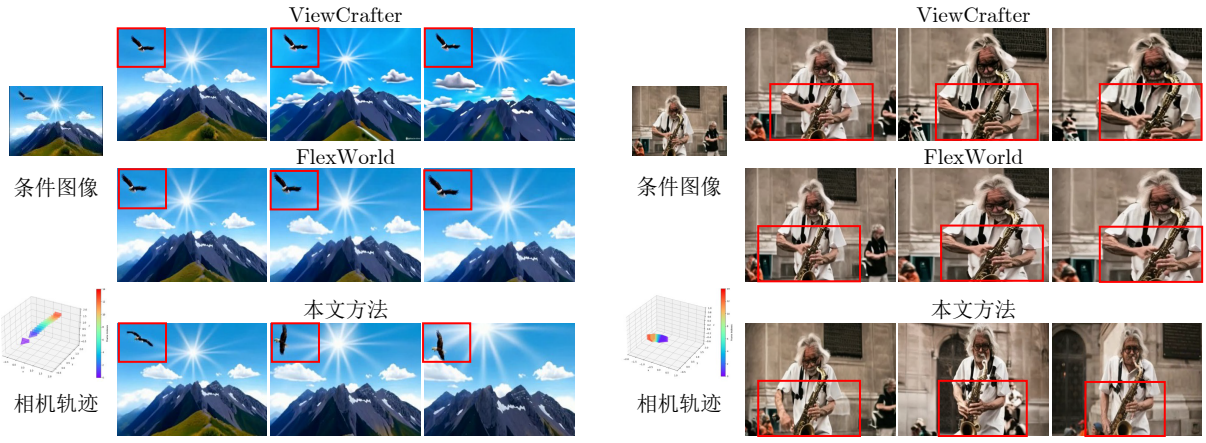

图 6　本文方法与 3D 新视角场景生成方法的比较

Fig. 6　Comparison between our method and 3D novel-view scene generation methods

表 2 消融实验
Table 2 Ablation experiment

方法	FID ↓	FVD ↓	COLMAP error rate ↓	Transition error ↓	Rotation error ↓
基线方法	11.12	72.31	18.9%	5.4319	2.1482
+ Inj.	9.97	61.13	7.6%	3.8512	1.9942
+ Inj. + Init.	9.85	58.39	7.1%	3.7125	1.9854
+ Inj. + Guid.	9.87	57.83	6.7%	3.7324	1.9691
本文方法	**9.61**	**52.63**	**5.2%**	**3.6372**	**1.9633**

合以上结论可知, 本文所提出的三种策略均能有效提高生成视频的视觉质量, 并引导模型生成更具空间一致性的视频内容. 此外, 虽然本文方法的推理用时有所提升, 即在单张 A800 GPU 上的推理用时由基线方法的 32 s 增至 121 s, 但考虑到生成视频质量的突破, 当前约 2 min 的推理时间在实际应用场景中仍处于可接受范围内.

4 结论

为解决相机轨迹可控的视频生成任务中的 3D 空间结构一致性不足问题, 本文提出利用基于深度的视角形变映射方法来为模型引入空间先验. 这一空间先验可在训练、推理两个阶段被引入模型. 在训练阶段, 本文设计 Warp-Injection, 通过逐帧视角形变与图像补全构建具备高度 3D 一致性的参考帧, 为模型提供空间结构先验; 在推理阶段, 本文提出 Warp-Init, 通过迭代式视角形变赋予初始化噪声优良的空间结构, 并结合 Warp-Guidance 在去噪过程中校正生成帧与预期目标视频的差异. 实验结果表明, 本文方法在所有评测指标上均显著优于现有的基线方法, 能够生成 3D 空间结构一致性更优的视频.

参考文献

1 Zhang L M, Rao A Y, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 3836−3847

2 Mou C, Wang X T, Xie L B, Wu Y Z, Zhang J, Qi Z G, et al. T2I-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024. 4296−4304

3 Wang C, Gu J X, Hu P W, Zhao H Y, Guo Y F, Han J H, et al. EasyControl: Transfer ControlNet to video diffusion for controllable generation and interpolation. arXiv preprint arXiv: 2408.13005, 2024.

4 Zhang Y B, Wei Y X, Jiang D S, Zhang X P, Zuo W M, Tian Q. ControlVideo: Training-free controllable text-to-video generation. arXiv preprint arXiv: 2305.13077, 2023.

5 Zhao M, Wang R Z, Bao F, Li C X, Zhu J. ControlVideo: Conditional control for one-shot text-driven video editing and beyond. arXiv preprint arXiv: 2305.17098, 2023.

6 Zhong Y, Zhao M, You Z B, Yu X F, Zhang C W, Li C X. PoseCrafter: One-shot personalized video synthesis following flexible pose control. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2025. 243−260

7 Zhu S H, Chen J L, Dai Z Z, Dong Z L, Xu Y H, Cao X, et al. Champ: Controllable and consistent human image animation with 3D parametric guidance. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2025. 145−162

8 Wang Z X, Yuan Z Y, Wang X T, Li Y W, Chen T S, Xia M H, et al. MotionCtrl: A unified and flexible motion controller for video generation. In: Proceedings of the ACM SIGGRAPH Conference Papers. Denver, USA: ACM, 2024. Article No. 114

9 He H, Xu Y H, Guo Y W, Wetzstein G, Dai B, Li H S, et al. CameraCtrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv: 2404.02101, 2024.

10 Xu D J, Nie W L, Liu C, Liu S F, Kautz J, Wang Z Y, et al. CamCo: Camera-controllable 3D-consistent image-to-video generation. arXiv preprint arXiv: 2406.02509, 2024.

11 Zheng G C, Li T, Jiang R, Lu Y H, Wu T, Li X. CamI2V: Camera-controlled image-to-video diffusion model. arXiv preprint arXiv: 2410.15957, 2024.

12 Hou C, Chen Z B. Training-free camera control for video generation. arXiv preprint arXiv: 2406.10126, 2024.

13 You M, Zhu Z Y, Liu H, Hou J H. NVS-Solver: Video diffusion model as zero-shot novel view synthesizer. arXiv preprint arXiv: 2405.15364, 2024.

14 Zhou T H, Tucker R, Flynn J, Fyffe G, Snavely N. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 2018, **37**(4): Article No. 65

15 Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 6629−6640

16 Unterthiner T, van Steenkiste S, Kurach K, Marinier R, Michalski M, Gelly S. Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv: 1812.01717, 2018.

17 Schönberger J L, Frahm J M. Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 4104−4113

18 Blattmann A, Rombach R, Ling H, Dockhorn T, Kim S W, Fidler S, et al. Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 22563−22575

19 Singer U, Polyak A, Hayes T, Yin X, An J, Zhang S Y, et al. Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv: 2209.14792, 2022.

20 Ho J, Chan W, Saharia C, Whang J, Gao R Q, Gritsenko A, et al. Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv: 2210.02303, 2022.

21 Guo Y W, Yang C Y, Rao A Y, Liang Z Y, Wang Y H, Qiao Y, et al. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv: 2307.04725, 2023.

22 Chen H X, Zhang Y, Cun X D, Xia M H, Wang X T, Weng C, et al. VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 7310−7320

23 Hong W Y, Ding M, Zheng W D, Liu X H, Tang J. CogVideo: Large-scale pretraining for text-to-video generation via Transformers. arXiv preprint arXiv: 2205.15868, 2022.

24 HaCohen Y, Chiprut N, Brazowski B, Shalem D, Moshe D, Richardson E, et al. LTX-Video: Realtime video latent diffusion. arXiv preprint arXiv: 2501.00103, 2024.

25 Yang Z Y, Teng J Y, Zheng W D, Ding M, Huang S Y, Xu J Z, et al. CogVideoX: Text-to-video diffusion models with an expert Transformer. arXiv preprint arXiv: 2408.06072, 2024.

26 Blattmann A, Dockhorn T, Kulal S, Mendelevitch D, Kilian M, Lorenz D, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv: 2311.15127, 2023.

27 Brooks T, Peebles B, Holmes C, DePue W, Guo Y F, Jing L, et al. Video generation models as world simulators [Online], available: https://openai.com/index/videogeneration-models-as-world-simulators, August 8, 2025

28 Bao F, Xiang C D, Yue G, He G D, Zhu H Z, Zheng K W, et al. Vidu: A highly consistent, dynamic and skilled text-to-video generator with diffusion models. arXiv preprint arXiv: 2405.04233, 2024

29 Xing J B, Xia M H, Zhang Y, Chen H X, Yu W B, Liu H Y, et al. DynamiCrafter: Animating open-domain images with video diffusion priors. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2025. 399−417

30 Zhang S W, Wang J Y, Zhang Y Y, Zhao K, Yuan H J, Qin Z W, et al. I2VGen-XL: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv: 2311.04145, 2023.

31 Chen X Y, Wang Y H, Zhang L J, Zhuang S B, Ma X, Yu J S, et al. SEINE: Short-to-long video diffusion model for generative transition and prediction. arXiv preprint arXiv: 2310.20700, 2023.

32 Zhao M, Zhu H Z, Xiang C D, Zheng K W, Li C X, Zhu J. Identifying and solving conditional image leakage in image-to-video diffusion model. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 954

33 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 574

34 Ho J, Salimans T. Classifier-free diffusion guidance. arXiv preprint arXiv: 2207.12598, 2022.

35 Zhao M, Bao F, Li C X, Zhu J. EGSDE: Unpaired image-to-image translation via energy-guided stochastic differential equations. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 261

36 Bao F, Zhao M, Hao Z K, Li P Y, Li C X, Zhu J. Equivariant energy-guided SDE for inverse molecular design. arXiv preprint arXiv: 2209.15408, 2023.

37 Wang X, Yuan H J, Zhang S W, Chen D Y, Wang J N, Zhang Y Y, et al. VideoComposer: Compositional video synthesis with motion controllability. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 334

38 Hu E J, Shen Y L, Wallis P, Allen-Zhu Z, Li Y Z, Wang S A, et al. LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv: 2106.09685, 2021.

39 Fehn C. Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. In: Proceedings of the SPIE 5291, Stereoscopic Displays and Virtual Reality Systems XI. San Jose, USA: SPIE, 2004. 93−104

40 Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 10684−10695

41 Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR, 2021. 8748−8763

42 Khachatryan L, Movsisyan A, Tadevosyan V, Henschel R, Wang Z Y, Navasardyan S, et al. Text2Video-Zero: Text-to-image diffusion models are zero-shot video generators. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 15954−15964

43 Ren W M, Yang H, Zhang G, Wei C, Du X R, Huang W H, et al. Consisti2V: Enhancing visual consistency for image-to-video generation. arXiv preprint arXiv: 2402.04324, 2024.

44 Ge S W, Nah S, Liu G L, Poon T, Tao A, Catanzaro B, et al. Preserve your own correlation: A noise prior for video diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 22930−22941

45 Burgert R, Xu Y C, Xian W Q, Pilarski O, Clausen P, He M M, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. arXiv preprint arXiv: 2501.08331, 2025.

46 Meng C L, He Y T, Song Y, Song J M, Wu J J, Zhu J Y, et al. SDEdit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv: 2108.01073, 2021.

47 Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 672

48 Yang L H, Kang B Y, Huang Z L, Zhao Z, Xu X G, Feng J S, et al. Depth anything V2. arXiv preprint arXiv: 2406.09414, 2024.

49 Black Forest Labs. Flux [Online], available: https://github.

com/black-forest-labs/flux, October 13, 2025

50 Chen L X, Zhou Z H, Zhao M, Wang Y K, Zhang G, Huang W H, et al. FlexWorld: Progressively expanding 3D scenes for flexiable-view synthesis. arXiv preprint arXiv: 2503.13265, 2025.

51 Yu W B, Xing J B, Yuan L, Hu W B, Li X Y, Huang Z P, et al. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv: 2409.02048, 2024.

52 Ren X C, Shen T C, Huang J H, Ling H, Lu Y F, Nimier-David M, et al. Gen3C: 3D-informed world-consistent video generation with precise camera control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 6121−6132

朱泓舟 主要研究方向为机器学习, 生成模型.
E-mail: zhuhz22@mails.thu.edu.cn
(**ZHU Hong-Zhou** His research interests include machine learning and generative models.)

杨 雪 主要研究方向为深度学习, 生成模型.
E-mail: squarl4111@gmail.com
(**YANG Xue** Her research interests include deep learning and generative models.)

赵 敏 清华大学计算机科学与技术系助理研究员. 主要研究方向为视觉内容生成.
E-mail: gracezhao1997@gmail.com
(**ZHAO Min** Assistant researcher in the Department of Computer Science and Technology, Tsinghua University. Her main research interest is visual content generation.)

李崇轩 中国人民大学高瓴人工智能学院副教授. 主要研究方向为生成模型. E-mail: chongxuanli@ruc.edu.cn
(**LI Chong-Xuan** Associate professor at the Gaoling School of Artificial Intelligence, Renmin University of China. His main research interest is generative models.)

朱 军 清华大学计算机科学与技术系教授. 主要研究方向为机器学习. 本文通信作者.
E-mail: dcszj@mail.tsinghua.edu.cn
(**ZHU Jun** Professor in the Department of Computer Science and Technology, Tsinghua University. His main research interest is machine learning. Corresponding author of this paper.)