摘要

文本驱动全景视频生成旨在根据给定文本描述，直接合成覆盖水平 360°、垂直 180° 的球面视场且时序连贯、空间一致的视频内容，是沉浸式媒体创作、虚拟现实交互、数字世界构建以及具身智能环境模拟中的关键基础能力。与普通视角视频生成相比，该任务不仅要求模型准确理解文本语义并生成合理运动，还要求其在球面空间中维持全局视觉关系的一致性。因此，仅把普通视角视频横向拉宽，并不能得到合格的全景视频。由于全景视频通常采用等距矩形投影表示，球面信号映射到二维平面后会表现出显著的尺度非均匀性，并在水平方向上具有天然的周期连续性要求，模型需要在文本描述引导下同时处理球面几何、时间动态和经度边界连续性等问题。

围绕这一任务，现有研究仍受到数据基础、模型适配和应用验证三方面限制。首先，高质量文本–全景视频配对数据不足，使模型难以稳定学习文本语义、动态变化和球面结构之间的对应关系。其次，普通视角视频生成模型缺乏对全景空间的内在建模能力；具体而言，等距矩形投影会在高纬区域，即靠近球面两极的位置，引入明显的几何拉伸，而普通视角视频生成模型又缺乏对左右边界周期连续性的内在建模能力，使得现有方法在高纬畸变、边界接缝和内容多样性方面仍存在显著局限。最后，现有工作对长时序生成、超分辨率、语义修补和视频外延等下游应用的验证仍不充分，难以全面说明模型是否具备可迁移的球面空间表示能力。

针对上述挑战，本文系统研究文本驱动全景视频生成中的数据资源构建、预训练能力迁移、球面空间适配和下游应用验证问题。首先，为支撑模型学习从普通视角视频到全景视频的空间表示迁移，本文整合多个公开全景视频数据源，建立自动化筛选、标注、质量控制与语义均衡流程，构建了高质量、带文本标注且场景多样的全景视频数据集 PanoVid。其次，本文提出全景视频生成框架 PanoWan，在尽量保留预训练文本生成视频模型生成先验的基础上，通过轻量化适配模块将普通视角视频生成能力迁移到全景域。进一步地，本文围绕全景视频的纬度尺度变化和经度周期连续性，设计纬度感知采样、旋转语义去噪和填充式像素解码等关键策略：纬度感知采样使初始噪声分布更符合球面频率特性，从而缓解高纬区域的结构畸变；旋转语义去噪与填充式像素解码分别从去噪过程和解码过程抑制左右边界的语义断裂与像素接缝。最后，本文通过定量评价、主观偏好、消融实验和多类下游应用验证方法的有效性与泛化能力。

本文的主要贡献如下：

1. 构建了面向文本驱动全景视频生成任务的高质量数据集 PanoVid，通过多源数据整合、自动化筛选标注和语义均衡处理，为模型训练、评测与下游扩展提供了可靠

的数据基础。

2. 提出了基于预训练文本生成视频模型的全景视频生成框架 PanoWan，在保留普通视角视频生成先验的同时，以较低迁移成本实现了向全景视频生成任务的有效适配，并在多项评价指标和下游任务中表现出较强性能。

3. 设计了面向球面空间特性的纬度/经度感知机制，包括用于缓解高纬畸变的纬度感知采样，以及用于实现经度边界连续的旋转语义去噪和填充式像素解码，从结构层面改善了全景视频生成中的几何失真与边界接缝问题。

总体而言，本文从全景视频与普通视角视频之间的球面空间表示差异出发，将高质量数据构建、预训练视频生成先验迁移和轻量化球面结构适配有机结合，为文本驱动全景视频生成提供了可行方案，也为复杂空间场景中的生成建模与沉浸式内容自动创作提供了研究基础。实验结果表明，PanoWan 在文本驱动全景视频生成任务上取得了更优的综合性能，生成结果在语义一致性、时空连贯性、高纬区域几何合理性与边界自然性方面均有明显提升，并能够零样本迁移到长时序生成、超分辨率、语义编辑和视频外延等多种下游场景，体现出良好的泛化能力与应用潜力。

关键词：全景视频, 视频生成, 扩散模型

Panoramic Video Generation Based on Latitude and Longitude-aware Mechanisms

Yifei Xia (Computer Application Technology)
Directed by: Prof. Boxin Shi

ABSTRACT

Text-driven panoramic video generation aims to create videos covering 360° horizontal and 180° vertical spherical views, while maintaining the temporal and spatial consistency of given text descriptions. It is a key capability for immersive media creation, virtual reality interaction, digital world building, and embodied environment simulation. Compared with conventional perspective video generation, this task not only requires accurate understanding of textual semantics and plausible motion synthesis, but also requires maintaining globally consistent visual relationships in the spherical space. Therefore, simply widening a conventional video cannot yield a valid panoramic video. Since panoramic videos are usually represented by equirectangular projection, the spherical signal shows non-uniform scale when mapped to a two-dimensional plane, and requires periodic continuity along the horizontal direction. Therefore, the model must handle spherical geometry, temporal dynamics and continuity across longitude edges under the guidance of text.

The existing research still has limitations in data resources, model adaptation and application-oriented validation. First, high-quality paired text-panoramic video data is still insufficient, which makes it difficult for the model to learn the stable matching between text semantics, temporal dynamics and spherical structure. Second, the conventional perspective video generation model lacks the inherent modeling of panoramic space. Especially, the equirectangular projection introduces strong geometric distortion in the high latitude region near the pole, while the conventional video generation model lacks the inherent mechanism of periodic continuity through the left and right edges. Therefore, the existing methods still show obvious limitations in high latitude distortion, seam suppression and content diversity. Third, the existing work provides limited validation for downstream tasks such as long video generation, super-resolution, semantic editing, and video outpainting, making it difficult to fully evaluate whether the learned representation can be transferred between panoramic applications.

To address these challenges, this thesis systematically studies the construction of data resources, the ability transfer of pretraining, the adaptation to spherical space, and the verification of downstream applications. Firstly, to support learning spatial representation transfer from conventional perspective videos to panoramic videos, we combine several public panoramic video sources and build an automated pipeline for filtering, annotation, quality control and semantic balance. This gives PANOVID, a high-quality panoramic video data set with text annotations and various scenes. Secondly, we propose PanoWan, a panoramic video generation framework that transfers the generative capability of pretrained text-to-video models to the panoramic domain with lightweight adaptation modules while preserving their strong priors as much as possible. Furthermore, we design latitude-aware sampling, rotated semantic denoising, and padded pixel-wise decoding to address latitudinal scale variation and longitudinal periodic continuity. Latitude-aware sampling aligns the initial noise distribution with spherical frequency properties to alleviate structural distortion in high-latitude regions, while rotated semantic denoising and padded pixel-wise decoding suppress semantic discontinuities and pixel-level seams at left-right boundaries during denoising and decoding, respectively. Finally, we validate the method through quantitative evaluation, user preference studies, ablation analysis, and multiple downstream applications.

The main contributions of this thesis are as follows:

1. We construct PANOVID, a high-quality dataset tailored to text-driven panoramic video generation, providing reliable data support for model training, evaluation, and downstream applications through multi-source data integration, automated filtering and annotation, and semantic balance.

2. We propose PanoWan, a panoramic video generation framework based on a pretrained text-to-video model, which effectively adapts conventional perspective video generation priors to panoramic video generation with low transfer cost and demonstrates strong performance across multiple metrics and downstream tasks.

3. We design latitude- and longitude-aware mechanisms for spherical spatial properties, including latitude-aware sampling for alleviating high-latitude distortion, as well as rotated semantic denoising and padded pixel-wise decoding for longitude boundary continuity, thereby improving geometric fidelity and seam suppression in panoramic video generation.

Overall, this thesis starts from the spherical representation gap between panoramic videos and conventional perspective videos, and combines high-quality data construction, pretrained video prior transfer, and lightweight spherical structural adaptation into a practical solution

for text-driven panoramic video generation. It also provides useful support for generative modeling in complex spatial scenes and automatic immersive content creation. Experimental results show that PanoWan achieves stronger overall performance on text-driven panoramic video generation, producing results with better semantic alignment, stronger spatio-temporal coherence, more reasonable high-latitude geometry, and more natural boundary transitions. It also generalizes in a zero-shot manner to downstream tasks such as long video generation, super-resolution, semantic editing, and video outpainting.

KEY WORDS: Panoramic Video, Video Generation, Diffusion Models