Hierarchical Video Generation: A Unified Architecture of Diffusion Models, Flow Matching, and Scene Language guided Semantic Control

Ayush Bodade
ayushbodade1@gmail.com

03 January 2024

Abstract

This document presents a research proposal focused on controlled video generation, utilizing diffusion models and flow matching combined leveraging Scene Language for architectural improvements, aiming to produce high-quality, coherent and controllable outputs.

Contents

Abstract

This proposal outlines an architecture for controlled video generation integrating diffusion models, flow matching, and Scene Language. The architecture leverages Scene Language for semantic planning, motion-guided flow matching for temporal coherence, hierarchical space-time diffusion for frame synthesis, and attention-based architectures for improved flow consistency. In addition, advanced generative mechanisms such as latent cross-frame refinement and style transfer are introduced. This approach aims to enhance controlled video generation by achieving high-quality, temporally consistent, and semantically controllable output.

Introduction

The generation of high-quality and temporally coherent videos remains a challenging task in generative modeling. While recent advancements have been made in the controlled video generation space, they often lack fine-grained control over narrative elements and scene composition. This proposal introduces a multifaceted approach integrating:

- **Diffusion Models**: For high-fidelity video frame synthesis.
- **Flow Matching**: To ensure realistic motion dynamics and temporal coherence.
- **Scene Language**: For dynamic scene planning and semantic control.
- **Hierarchical Generative Structures**: To capture multi-scale temporal and spatial relationships.
- **Cross-frame Latent Refinement**: To maintain consistency across sequences.
- **Style Transfer and Fine-grained Control**: Using contextual information and user guidance.

Architecture

Temporal Segmentation and Frame Generation: Define the total duration of the video T and the frame rate f. Segment the timeline into N time blocks, each of length $T_{\text{block}} = T/N$. Each block corresponds to a sequence of frames:

$$F_t = \{F_{t,1}, F_{t,2}, \ldots, F_{t,m}\},$$

where $m = T_{\text{block}} \times f$.

Dynamic Scene Planning using Scene Language: For each time block t, an LLM generates a Dynamic Scene Syntax (DSS) that encapsulates high-level descriptions, object interactions, and motion patterns:

$$\text{DSS}_t = \{D_t^s, L_t^o, M_t^b\}$$

where:

- D_t^s: High-level scene description.
- L_t^o: Object layouts and interactions.
- M_t^b: Background motion patterns.

The LLM iteratively refines the DSS based on feedback, ensuring semantic alignment with generated frames. This phase of the architecture can make use of Scene Langauge, as described in the previous reports. Time blocks can be divided into furthur chunks to gain more control.

Hierarchical Space-Time Diffusion Model: A Hierarchical Space-Time U-Net (HST-UNet) synthesizes frames for each block. Let $z_t^{(0)}$ represent the initial noisy latent variable, initialized based on DSS:

$$z_t^{(k-1)} = f_\theta(z_t^{(k)}, \text{DSS}_t),$$

for $k = 1, \ldots, K$, where f_θ is the denoising function conditioned on DSS.

Hierarchical Latent Refinement: To improve fidelity and coherence, we use hierarchical latent spaces at multiple scales:

$$z_t^{(0)} = \{z_t^{(0,1)}, z_t^{(0,2)}, z_t^{(0,3)}\},$$

where each level focuses on capturing details at different levels of abstraction (e.g. global, regional, local).

where self-attention and cross-attention are applied across different scales to preserve temporal consistency.

Flow Matching for Temporal Coherence: Flow matching improves temporal consistency by estimating motion fields:

$$v_t(x) = \nabla \log p_t(x),$$

and minimizing:

$$L_{\text{flow}} = \mathbb{E}_{t,x} \| v_t(x) - \hat{v}_t(x) \|^2$$

where $\hat{v}_t(x)$ is the predicted motion field.

Flow Matching with Pyramidal Representations: Employ a pyramidal flow matching algorithm that operates across multiple spatial and temporal resolutions to efficiently model complex motions and maintain temporal consistency.

- **Spatial Pyramid:** Process frames at progressively higher resolutions to capture fine details.
- **Temporal Pyramid:** Compress and process sequences at varying temporal resolutions to manage long-range dependencies.

Cross-frame Latent Refinement: Refining latent variables between frames allows for better motion consistency:

$$z_t^{(k)} = z_t^{(k-1)} + f_\phi(z_t^{(k-1)}, z_{t+1}^{(k-1)})$$

where f_ϕ is a refinement function that captures the coherence of the cross-frame.

Temporal Attention Mechanisms: Use temporal attention to improve coherence by aggregating information across frames:

$$F_{t,i} = \text{Attention}(F_{t,i}, \{F_{t-1,m}, F_{t,i-1}\})$$

with mechanisms such as Longformer to capture long-range dependencies. Flex/Flash attention can be used between the time blocks and the pages of each time block.

Style Transfer and Fine-grained Control: Leveraging LLM-guided style transfer, we enhance content consistency and user-defined attributes. For each frame:

$$F_{t,i} = \text{StyleTransfer}(F_{t,i}, \text{DSS}_t)$$

where user-specific style embeddings are integrated into the diffusion process.

Motion-Guided Noise Initialization: Initialize the noise $z_t^{(0)}$ using the motion dynamics derived from DSS:

$$z_t^{(0)} \sim \mathcal{N}(0, \Sigma(M_t^b)),$$

where $\Sigma(M_t^b)$ encodes motion patterns.

Video Generation Workflow

The exact flow for generating the video is as follows:

1. **Input Specification:** Define video parameters such as total duration T, frame rate f, and segment the timeline into N time blocks of size T_{block}.
2. **Dynamic Scene Planning:** For each time block t, generate a Dynamic Scene Syntax (DSS) using the LLM, describing the scene, object layouts, and motion patterns.
3. **Noise Initialization:** Initialize the noise $z_t^{(0)}$ for the diffusion process based on motion dynamics from the DSS.
4. **Frame Synthesis:** Use the Hierarchical Space-Time Diffusion Model (HST-UNet) to synthesize frames for the time block by iteratively denoising $z_t^{(k)}$ over K steps, conditioned on the DSS.
5. **Flow Matching:** Enhance motion consistency by applying flow matching between latent states of adjacent frames, aligning them to predicted motion fields.

6. **Temporal Attention:** Flow of cross-frame information using attention mechanisms to ensure smooth transitions across frames, pages and blocks.
7. **Cross-frame Latent Refinement:** Refine latent variables using historical frame information to ensure consistency.
8. **Style Transfer and Control:** Incorporate user-guided style transfer and attribute control to enhance perceptual quality and coherence.
9. **Block Stitching:** Combine frames from consecutive time blocks by smoothing the transition between the last frame of block t and the first frame of block $t+1$.
10. **Output Assembly:** Assemble the synthesized frames into the final video sequence, ensuring temporal coherence and high-quality rendering.

Training Objectives

The model optimizes a joint loss:

$$L = L_{\text{diffusion}} + \lambda_1 L_{\text{flow}} + \lambda_2 L_{\text{attention}} + \lambda_3 L_{\text{refinement}} + \lambda_4 L_{\text{style}},$$

where:

- $L_{\text{diffusion}}$: Standard diffusion loss for denoising.
- L_{flow}: Flow matching loss for motion consistency.
- $L_{\text{attention}}$: Temporal attention loss.
- $L_{\text{refinement}}$: Cross-frame latent refinement loss.
- L_{style}: Style transfer and control loss.

Evaluation Metrics

Evaluate the proposed model on:

- **Perceptual Quality**: Using FID and IS scores.
- **Temporal Coherence**: Measuring consistency across frames using motion metrics.
- **Control Accuracy**: Assessing alignment with user-provided text descriptions and style attributes.
- **Flow Consistency**: Assessing the accuracy of flow fields and their alignment with ground-truth motion dynamics.

References

[1] Omer Bar-Tal, Hila Chefer, Omer Tov, et al. Lumiere: A Space-Time Diffusion Model for Video Generation. 2024. Available at: https://arxiv.org/abs/2401.12945.

[2] Daniel Watson, Saurabh Saxena, Lala Li, et al. Controlling Space and Time with Diffusion Models. 2024. Available at: https://arxiv.org/abs/2407.07860.

[3] Yang Jin, Zhicheng Sun, Ningyuan Li, et al. Pyramidal Flow Matching for Efficient Video Generative Modeling. 2024. Available at: https://arxiv.org/abs/2410.05954.

[4] Yunzhi Zhang, Zizhang Li, Matt Zhou, et al. The Scene Language: Representing Scenes with Programs, Words, and Embeddings. 2024. Available at: https://arxiv.org/abs/2410.16770.

[5] Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, et al. Adjoint Matching: Fine-tuning Flow and Diffusion Generative Models with Memoryless Stochastic Optimal Control. 2024. Available at: https://arxiv.org/abs/2409.08861.

[6] Kunpeng Song, Tingbo Hou, Zecheng He, et al. DirectorLLM for Human-Centric Video Generation. 2024. Available at: https://arxiv.org/abs/2412.14484.

[7] Tianyi Zhu, Dongwei Ren, Qilong Wang, et al. Generative Inbetweening through Frame-wise Conditions-Driven Video Generation. 2024. Available at: https://arxiv.org/abs/2412.11755.

[8] Haitao Zhou, Chuang Wang, Rui Nie, et al. TrackGo: A Flexible and Efficient Method for Controllable Video Generation. 2024. Available at: https://arxiv.org/abs/2408.11475.