Novelty Assessment Report

Paper: CineTrans: Learning to Generate Videos with Cinematic Transitions via Masked Diffusion Models
PDF URL: https://openreview.net/pdf?id=955hVLJdfP
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Despite significant advances in video synthesis, research into multi-shot video generation remains in its infancy. Even with scaled-up models and massive datasets, the shot transition capabilities remain rudimentary and unstable, largely confining generated videos to single-shot sequences. In this work, we introduce CineTrans, a novel framework for generating coherent multi-shot videos with cinematic, film-style transitions. To facilitate insights into the film editing style, we construct a multi-shot video-text dataset Cine250K with detailed shot annotations. Furthermore, our analysis of existing video diffusion models uncovers a correspondence between attention maps in the diffusion model and shot boundaries, which we leverage to design a mask-based control mechanism that enables transitions at arbitrary positions and transfers effectively in a training-free setting. After fine-tuning on our dataset with the mask mechanism, CineTrans produces cinematic multi-shot sequences while adhering to the film editing style, avoiding unstable transitions or naive concatenations. Finally, we propose specialized evaluation metrics for transition control, temporal consistency and overall quality, and demonstrate through extensive experiments that CineTrans significantly outperforms existing baselines across all criteria.

Core Task Landscape

This paper addresses: **Generating Multi-Shot Videos with Cinematic Transitions**
A total of **40 papers** were analyzed and organized into a taxonomy with **11 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Multi-Shot Video Generation Frameworks**
- **Cinematic Transition and Editing Control**
- **Camera and Shot Control for Video Generation**
- **Datasets and Benchmarks for Cinematic Video Generation**
- **Domain-Specific and Multimodal Cinematic Video Synthesis**
- **Supporting Methods for Video Synthesis**

Complete Taxonomy Tree

- Generating Multi-Shot Videos with Cinematic Transitions Survey Taxonomy
- Multi-Shot Video Generation Frameworks
  - Narrative-Driven Multi-Shot Synthesis (5 papers)
  - [1] VideoGen-of-Thought: Step-by-step generating multi-shot video with minimal manual intervention (Zheng Mingzhe, 2024) View paper
  - [8] Holocine: Holistic generation of cinematic multi-shot long video narratives (Ouyang Hao, 2025) View paper
  - [16] Text2Story: Advancing Video Storytelling with Text Guidance (Kang, 2025) View paper
  - [17] MovieFactory: Automatic Movie Creation from Text using Large Generative Models for Language and Images (Jun-Chen Zhu, 2023) View paper
  - [37] VideoGen-of-Thought: A Collaborative Framework for Multi-Shot Video Generation (Zheng Mingzhe, 2024) View paper
  - Keyframe-Anchored Multi-Shot Generation (3 papers)
  - [2] CineVerse: Consistent Keyframe Synthesis for Cinematic Scene Composition (Phung, 2025) View paper
  - [14] STORYANCHORS: Generating Consistent Multi-Scene Story Frames for Long-Form Narratives (Wang Bo, 2025) View paper
  - [24] STAGE: Storyboard-Anchored Generation for Cinematic Multi-shot Narrative (Peixuan Zhang, 2025) View paper
  - Attention-Based Multi-Shot Control (3 papers)
  - [3] ShotAdapter: Text-to-Multi-Shot Video Generation with Diffusion Models (Ozgur Kara, 2025) View paper
  - [13] DiTCtrl: Exploring Attention Control in Multi-Modal Diffusion Transformer for Tuning-Free Multi-Prompt Longer Video Generation (Ming-Hong Cai, 2024) View paper
  - [25] MultiShotMaster: A Controllable Multi-Shot Video Generation Framework (Qinghe Wang, 2025) View paper
- Cinematic Transition and Editing Control
  - Transition-Aware Video Synthesis ★ (4 papers)
  - [0] CineTrans: Learning to Generate Videos with Cinematic Transitions via Masked Diffusion Models (Anon et al., 2026) View paper
  - [5] StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text (Roberto Henschel, 2024) View paper
  - [11] MAVIN: Multi-Action Video Generation with Diffusion Models via Transition Video Infilling (Zhang Bo-wen, 2024) View paper
  - [15] ShotDirector: Directorially Controllable Multi-Shot Video Generation with Cinematographic Transitions (Xiaoxue Wu, 2025) View paper
  - Editing Pattern and Cinematographic Language Learning (5 papers)
  - [4] Shot sequence ordering for video editing: Benchmarks, metrics, and cinematology-inspired computing methods (Li Yuzhi, 2025) View paper

- [7] Can video generation replace cinematographers? Research on the cinematic language of generated video (Li Xiaozhe, 2024) View paper
- [10] FilMaster: Bridging Cinematic Principles and Generative AI for Automated Film Generation (Huang, 2025) View paper
- [12] Cut2Next: Generating Next Shot via In-Context Tuning (He Jingwen, 2025) View paper
- [26] Cinematographic-Aware Coherent Shot Assembly for How-To Vlog Generation (Yuqi Zhang, 2025) View paper

• Camera and Shot Control for Video Generation
- Parametric Camera Trajectory Control (3 papers)
- [6] MotionCanvas: Cinematic Shot Design with Controllable Image-to-Video Generation (Xing, 2025) View paper
- [20] CineLOG: A Training Free Approach for Cinematic Long Video Generation (Zahra Dehghanian, 2025) View paper
- [21] Infinite Video Generation with Cinematic Camera Trajectory Control (J Lee, 2025) View paper
- Virtual Cinematography and Shot Planning (2 papers)
- [18] Dynamic Storyboard Generation in an Engine-based Virtual Environment for Video Production (Anyi Rao, 2023) View paper
- [38] JAWS: Just A Wild Shot for Cinematic Transfer in Neural Radiance Fields (Xi Wang, 2023) View paper

• Datasets and Benchmarks for Cinematic Video Generation (2 papers)
- [9] TalkCuts: A Large-Scale Dataset for Multi-Shot Human Speech Video Generation (Chen, 2025) View paper
- [31] LoCoT2V-Bench: A Benchmark for Long-Form and Complex Text-to-Video Generation (Zheng Xiangqing, 2025) View paper

• Domain-Specific and Multimodal Cinematic Video Synthesis
- Audio-Driven and Multimodal Cinematic Synthesis (3 papers)
- [27] Multimodal Cinematic Video Synthesis Using Text-to-Image and Audio Generation Models (S. Sridhar, 2025) View paper
- [28] Wan-S2V: Audio-Driven Cinematic Video Generation (Gao Xin, 2025) View paper
- [40] JenBridge: Adaptive Long-Form Video Soundtracking across Scene Transition (J Yu, n.d.) View paper
- Specialized Production and Editing Applications (3 papers)
- [22] Fine-Tuning Open Video Generators for Cinematic Scene Synthesis: A Small-Data Pipeline with LoRA and Wan2.1 I2V (Meftun Akarsu, 2025) View paper
- [23] EditIQ: Automated Cinematic Editing of Static Wide-Angle Videos via Dialogue Interpretation and Saliency Cues (Rohit Girmaji, 2025) View paper
- [39] GAZED- Gaze-guided Cinematic Editing of Wide-Angle Monocular Video Recordings (K. L. Bhanu Moorthy, 2022) View paper

• Supporting Methods for Video Synthesis (8 papers)
- [19] Trans4D: Realistic Geometry-Aware Transition for Compositional Text-to-4D Synthesis (Zeng, 2024) View paper
- [29] TeViS: Translating Text Synopses to Video Storyboards (Xu Gu, 2022) View paper
- [30] Multi-Conditional Generative Adversarial Network for Text-to-Video Synthesis (Rui Zhou, 2022) View paper
- [32] Artificial Intelligence in Multimedia Content Generation: A Review of Audio and Video Synthesis Techniques (C Ding, 2025) View paper
- [33] From Loop to Video Essay (ATTRACTIONS, 2025) View paper
- [34] TimeChat-Online: 80% Visual Tokens are Naturally Redundant in Streaming Videos (Linli Yao, 2025) View paper
- [35] MultiCOIN: Multi-Modal COntrollable Video INbetweening (Tanveer, 2025) View paper
- [36] Fan Video of Attractions: From Loop to Video Essay (Stein, 2025) View paper

Narrative

Core task: generating multi-shot videos with cinematic transitions. This emerging field addresses the challenge of synthesizing coherent video sequences that span multiple shots with professional-quality transitions and editing conventions. The taxonomy reveals several complementary research directions: Multi-Shot Video Generation Frameworks develop end-to-end systems for producing complete video narratives; Cinematic Transition and Editing Control focuses on modeling shot boundaries, cuts, and temporal coherence across scenes; Camera and Shot Control methods enable precise manipulation of cinematographic parameters like framing and movement; Datasets and Benchmarks establish evaluation standards for cinematic quality; Domain-Specific and Multimodal approaches tailor generation to particular content types or input modalities; and Supporting Methods provide foundational techniques for video synthesis. Works like VideoGen of Thought[1] and CineVerse[2] exemplify comprehensive frameworks, while ShotAdapter[3] and Shot Sequence Ordering[4] tackle specific aspects of shot-level control and sequencing.

Recent efforts reveal a tension between holistic narrative generation and fine-grained control over individual shots and transitions. Some approaches prioritize seamless temporal extension and transition smoothness, as seen in StreamingT2V[5] and related autoregressive methods, while others emphasize explicit modeling of cinematic grammar through shot planning and editing rules. CineTrans[0] sits within the transition-aware synthesis cluster, focusing specifically on learning and generating natural transitions between shots—a capability that distinguishes it from broader multi-shot frameworks like ShotDirector[15] or MAVIN[11], which may emphasize shot composition or narrative structure over transition quality. This positioning reflects a growing recognition that professional video synthesis requires not just generating individual shots but also mastering the subtle temporal and visual continuity that defines cinematic storytelling, an area where explicit transition modeling offers advantages over purely autoregressive or frame-by-frame generation strategies.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text

Authors: Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, et al. (9 authors total) | **Year/Venue**: 2024 • Computer Vision and Pattern Recognition | **URL**: View paper

Abstract

Text-to-video diffusion models enable the generation of high-quality videos that follow text instructions, simplifying the process of producing diverse and individual content. Current methods excel in generating short videos (up to 16s), but produce hard-cuts when naively extended to long video synthesis. To overcome these limitations, we present StreamingT2V, an autoregressive method that generates long videos of up to 2 minutes or longer with seamless transitions. The key components are: (i) a...

Relationship Analysis

Both papers belong to the Transition-Aware Video Synthesis category, focusing on generating videos with smooth transitions between shots using specialized mechanisms within diffusion models. StreamingT2V addresses long video generation through autoregressive chunk-based synthesis with conditional attention modules (CAM) and appearance preservation modules (APM) to maintain consistency across extended sequences, while CineTrans focuses specifically on cinematic multi-shot videos with film-style transitions using a mask-based attention mechanism that enforces block-diagonal attention patterns aligned with shot boundaries. The key difference is that

StreamingT2V targets seamless chunk concatenation for arbitrary-length videos, whereas CineTrans emphasizes precise frame-level control of cinematic transitions at user-specified positions informed by film editing conventions.

---

2. MAVIN: Multi-Action Video Generation with Diffusion Models via Transition Video Infilling

Authors: Zhang Bo-wen, Bowen Zhang, Xie, Xiaofei, Xiaofei Xie, et al. (16 authors total) | **Year/Venue**: 2024 • arXiv.org | **URL**: View paper

Abstract

Diffusion-based video generation has achieved significant progress, yet generating multiple actions that occur sequentially remains a formidable task. Directly generating a video with sequential actions can be extremely challenging due to the scarcity of fine-grained action annotations and the difficulty in establishing temporal semantic correspondences and maintaining long-term consistency. To tackle this, we propose an intuitive and straightforward solution: splicing multiple single-action vid...

Relationship Analysis

Both papers belong to the Transition-Aware Video Synthesis category, focusing on generating smooth transitions between video shots using specialized mechanisms within diffusion models. While CineTrans addresses cinematic multi-shot video generation by analyzing attention maps and introducing a mask mechanism to control transitions at arbitrary positions with film-editing style, MAVIN tackles the transition problem through a video infilling approach that generates intermediate transition videos to connect pre-existing single-action segments. The key difference lies in their approach: CineTrans generates entire multi-shot videos with embedded transitions in a single pass using attention masking, whereas MAVIN generates transitions separately by infilling gaps between independently created video segments.

---

3. ShotDirector: Directorially Controllable Multi-Shot Video Generation with Cinematographic Transitions

Authors: Xiaoxue Wu, Xinyuan Chen, Yaohui Wang, Yu Qiao | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Shot transitions play a pivotal role in multi-shot video generation, as they determine the overall narrative expression and the directorial design of visual storytelling. However, recent progress has primarily focused on low-level visual consistency across shots, neglecting how transitions are designed and how cinematographic language contributes to coherent narrative expression. This often leads to mere sequential shot changes without intentional film-editing patterns. To address this limitatio...

Relationship Analysis

Both papers belong to the Transition-Aware Video Synthesis category, focusing on explicitly modeling cinematic transitions in multi-shot video generation using diffusion models. They share overlapping approaches in using mask mechanisms to control shot transitions and constructing specialized datasets with shot annotations. The key differences are that CineTrans analyzes attention maps to design a block-diagonal mask mechanism and constructs the Cine250K dataset with frame-level shot labels, while ShotDirector integrates parameter-level camera control (6-DoF poses, Pl¨ucker embeddings) with hierarchical editing-pattern-aware prompting and constructs the ShotWeaver40K dataset with professional editing pattern annotations (cut-in, cut-out, shot/reverse-shot, multi-angle).

Contributions Analysis

Overall novelty summary. The paper introduces CineTrans, a framework for generating multi-shot videos with cinematic transitions, alongside the Cine250K dataset with shot annotations and a mask-based control mechanism. It resides in the 'Transition-Aware Video Synthesis' leaf, which contains only four papers total, including this one. This leaf sits within the broader 'Cinematic Transition and Editing Control' branch, indicating a relatively sparse research direction focused specifically on modeling shot boundaries and transition quality rather than general multi-shot generation or narrative structure.

The taxonomy reveals that neighboring leaves address related but distinct challenges: 'Editing Pattern and Cinematographic Language Learning' (five papers) focuses on learning professional editing conventions from film data, while 'Narrative-Driven Multi-Shot Synthesis' (five papers) emphasizes story decomposition and shot-by-shot generation. The 'Attention-Based Multi-Shot Control' leaf (three papers) explores attention mechanisms for cross-shot consistency. CineTrans bridges transition modeling with editing pattern learning by constructing an annotated dataset and leveraging attention map analysis, positioning it at the intersection of explicit transition control and data-driven cinematic language understanding.

Among 30 candidates examined, the Cine250K dataset contribution shows no clear refutation across 10 candidates, suggesting novelty in providing detailed shot annotations for multi-shot video generation. The mask-based control mechanism examined 10 candidates with 4 appearing to provide overlapping prior work, indicating more substantial existing research on attention-based or mask-based transition control. The CineTrans framework itself examined 10 candidates with 1 refutable match, suggesting the integrated system approach may offer incremental novelty over existing transition-aware methods within this limited search scope.

Based on the top-30 semantic matches examined, the work appears to contribute primarily through its dataset and integrated framework rather than fundamentally novel transition mechanisms. The sparse taxonomy leaf (four papers) suggests transition-aware synthesis remains an emerging direction, though the refutation statistics indicate that specific technical components overlap with prior attention-based control methods. The analysis does not cover exhaustive literature beyond these candidates, leaving open questions about broader field coverage.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Cine250K multi-shot video dataset with shot annotations

Description: The authors develop a dataset of 250K video-text pairs featuring frame-level shot labels and hierarchical captions. This dataset captures film editing style and provides prior knowledge for generating cinematic multi-shot sequences.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. MovieBench: A Hierarchical Movie Level Dataset for Long Video Generation

URL: View paper

Brief Assessment

MovieBench[49] focuses on movie-length videos with multi-scene narratives and character consistency, while the original paper's Cine250K targets cinematic transitions within shorter multi-shot sequences (8-15s). These serve different purposes in video generation research.

---

2. Procedure-Aware Surgical Video-language Pretraining with Hierarchical Knowledge Augmentation

URL: View paper

Brief Assessment

Procedure Aware Surgical[42] focuses on surgical video-language pretraining with procedure-level annotations, not cinematic multi-shot video generation with film editing transitions. The domains and objectives are fundamentally different.

---

3. FineBio: A Fine-Grained Video Dataset of Biological Experiments with Hierarchical Annotation

URL: View paper

Brief Assessment

FineBio[50] focuses on biological experiments with hierarchical annotation of protocols, steps, and atomic operations, not on cinematic multi-shot video generation with film editing style transitions.

---

4. Video paragraph captioning using hierarchical recurrent neural networks

URL: View paper

Brief Assessment

Video Paragraph Captioning[48] focuses on generating paragraph descriptions for videos using hierarchical RNNs, not on creating multi-shot video datasets with shot annotations and hierarchical captions for video generation tasks.

---

5. AnimeShooter: A Multi-Shot Animation Dataset for Reference-Guided Video Generation

URL: View paper

Brief Assessment

AnimeShooter[44] focuses on animation-specific multi-shot datasets with character reference images for consistent character guidance, while the original paper targets cinematic film-editing style with general video content. The datasets serve different domains and purposes.

---

6. SpatialVID: A Large-Scale Video Dataset with Spatial Annotations

URL: View paper

Brief Assessment

SpatialVID[41] focuses on spatial intelligence with camera poses and depth annotations for 3D vision tasks, not on multi-shot video generation with hierarchical captions and shot-level editing style.

---

7. HecVL: Hierarchical Video-Language Pretraining for Zero-shot Surgical Phase Recognition

URL: View paper

Brief Assessment

HecVL[43] focuses on surgical video-language pretraining with hierarchical text annotations for medical procedures, not general multi-shot cinematic video datasets with film editing style annotations.

---

8. Video recap: Recursive captioning of hour-long videos

URL: View paper

Brief Assessment

Video Recap[46] focuses on hierarchical video captioning for hour-long videos with recursive caption generation at multiple temporal granularities, not on multi-shot video generation with cinematic transitions. The dataset structures and objectives differ fundamentally from Cine250K's focus on film editing style and shot-level annotations.

---

9. LVD-2M: A Long-take Video Dataset with Temporally Dense Captions

URL: View paper

Brief Assessment

LVD-2M[45] focuses on long-take videos without cuts, while Cine250K specifically targets multi-shot videos with shot transitions and hierarchical captions for cinematic editing.

---

10. Hierarchical Video-Moment Retrieval and Step-Captioning

URL: View paper

Brief Assessment

Hierarchical Video Moment[47] focuses on instructional video retrieval with step-level annotations for question-answering tasks, not on cinematic multi-shot video generation with film editing style transitions. The datasets serve fundamentally different purposes and application domains.

Contribution 2: Mask-based control mechanism for cinematic transitions

Description: The authors introduce a block-diagonal mask mechanism applied to attention layers in diffusion models. This mechanism enforces strong intra-shot correlations and weak inter-shot correlations, enabling precise frame-level control of cinematic transitions even without fine-tuning.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Dreamix: Video diffusion models are general video editors

URL: View paper

Brief Assessment

Dreamix[55] focuses on video editing through noise-based corruption and finetuning, not on mask-based attention control for frame-level transitions. The candidate does not demonstrate prior work on block-diagonal mask mechanisms for cinematic shot transitions.

---

2. FreeTraj: Tuning-Free Trajectory Control in Video Diffusion Models

URL: View paper

Brief Assessment

FreeTraj[56] focuses on trajectory control for object motion paths in video generation, not cinematic shot transitions. The mask mechanisms serve fundamentally different purposes: FreeTraj[56] modifies noise sampling and attention for motion trajectory control, while the original paper uses block-diagonal masks to enforce shot boundaries and cinematic transitions between distinct scenes.

---

3. TRIP: Temporal Residual Learning with Image Noise Prior for Image-to-Video Diffusion Models

URL: View paper

Brief Assessment

TRIP[60] focuses on image-to-video generation using temporal residual learning and image noise priors, not on mask-based attention control for frame-level transitions in multi-shot video generation.

---

4. Diffusion action segmentation

URL: View paper

Brief Assessment

Diffusion Action Segmentation[59] applies masking to action segmentation in videos, not cinematic transitions in video generation. The candidate focuses on temporal action classification, while the original addresses frame-level transition control in diffusion-based video synthesis.

---

5. DiTCtrl: Exploring Attention Control in Multi-Modal Diffusion Transformer for Tuning-Free Multi-Prompt Longer Video Generation

URL: View paper

Prior Art Analysis

DiTCtrl[13] demonstrates that mask-based attention control for frame-level transitions in diffusion models was explored prior to the original paper. Both papers apply block-diagonal mask mechanisms to attention layers in diffusion transformers to control temporal transitions. DiTCtrl[13] explicitly describes using mask-guided semantic control across different prompts with attention sharing, achieving smooth transitions without fine-tuning. The original paper's claim of being first to introduce a block-diagonal mask mechanism for cinematic transitions is refuted by DiTCtrl[13]'s prior work on mask-based attention control for multi-prompt video generation with smooth transitions.

Evidence

Evidence 1 - **Rationale**: Both papers claim training-free mask-based control mechanisms. DiTCtrl[13] explicitly states it is the 'first time' for training-free multi-prompt generation under mm-dit architectures, predating the original paper's similar claim. - **Original**: we introduce a mask-based control mechanism that enables transitions at arbitrary positions and transfers effectively in a training-free setting - **Candidate**: we propose ditctrl, a training-free multi-prompt video generation method under mm-dit architectures for the first time

Evidence 2 - **Rationale**: Both papers describe mask mechanisms that control correlations between different segments (shots/prompts) in video generation. DiTCtrl[13] describes mask-guided control for transitions between prompts, which is conceptually equivalent to the original paper's shot transitions. - **Original**: we introduce a mask mechanism featuring strong correlations within shots and weak correlations between shots, achieving controlled cinematic transitions and enabling zero-shot multi-shot generation - **Candidate**: finding that the 3d full attention behaves similarly to that of the cross/self-attention blocks in the unet-like diffusion models, enabling mask-guided precise semantic control across different prompts with attention sharing for multi-prompt video generation

Evidence 3 - **Rationale**: Both papers achieve frame-level transition control without training. DiTCtrl[13]'s 'smooth transitions' with 'multiple sequential prompts without additional training' directly parallels the original paper's 'precise frame-level cinematic transitions' in a 'training-free setting'. - **Original**: the application of mask enables strong intra-shot frame correlations in attention module, facilitating precise frame-level cinematic transitions, which remains effective even in a training-free setting - **Candidate**: based on our careful design, the video generated by ditctrl achieves smooth transitions and consistent object motion given multiple sequential prompts without additional training

---

6. Mask^ 2DiT: Dual Mask-based Diffusion Transformer for Multi-Scene Long Video Generation

URL: View paper

Prior Art Analysis

Mask2DiT[57] demonstrates prior work on mask-based attention control for shot transitions in diffusion models. Both papers apply block-diagonal mask mechanisms to attention layers in diffusion transformers to control shot transitions. The candidate paper explicitly introduces 'a symmetric binary mask at each attention layer within the dit architecture' for multi-scene video generation, which directly parallels the original paper's block-diagonal mask mechanism for cinematic transitions. Both approaches enforce strong intra-shot correlations and weak inter-shot correlations through attention masking, and both demonstrate effectiveness in training-free settings.

Evidence

Evidence 1 - **Rationale**: Both papers demonstrate that mask mechanisms can be applied to pre-trained models for shot transition control. The original explicitly mentions 'training-free setting' effectiveness, while the candidate shows adaptation of 'pre-trained dit-based t2v model' using attention masks. - **Original**: the application of mask enables strong intra-shot frame correlations in attention module, facilitating precise frame-level cinematic transitions, which remains effective even in a training-free setting - **Candidate**: with the incorporation of this attention mask, we effectively adapt the pre-trained dit-based t2v model for video generation tasks involving a fixed number of scenes

Evidence 2 - **Rationale**: Both papers identify attention mechanisms as key to controlling shot transitions and propose mask-based solutions. The original observes 'strong connection between attention probabilities and shot transitions' leading to a mask mechanism, while the candidate addresses alignment issues through 'attention mask mattn.' - **Original**: we analyze attention maps in diffusion models for multi-shot video generation and observe a strong connection between attention probabilities and shot transitions. building on this insight, we introduce a mask mechanism that enables cinematic transitions within diffusion models - **Candidate**: however, straightforward token concatenation disrupts the fine-grained alignment between each scene and its corresponding text annotation. to address this, we propose introducing an attention mask mattn to restore the fine-grained, one-to-one alignment

---

7. EIDT-V: Exploiting Intersections in Diffusion Trajectories for Model-Agnostic, Zero-Shot, Training-Free Text-to-Video Generation

URL: View paper

Brief Assessment

EIDT V[61] focuses on grid-based prompt switching for image-to-video generation using diffusion trajectory intersections, not mask-based attention control for frame-level transitions in video diffusion models. The technical approaches are fundamentally different.

---

8. Peekaboo: Interactive video generation via masked-diffusion

URL: View paper

Brief Assessment

Peekaboo[58] focuses on interactive video generation with spatio-temporal control for object placement and movement, not cinematic shot transitions. The mask mechanism in Peekaboo[58] controls object locations within frames, while the original paper's contribution addresses shot-level transitions in multi-shot video generation.

---

9. ShotAdapter: Text-to-Multi-Shot Video Generation with Diffusion Models

URL: View paper

Prior Art Analysis

ShotAdapter[3] demonstrates prior work on mask-based attention control for multi-shot video generation with diffusion models. Both papers employ masking strategies applied to attention layers to control shot transitions, though with different specific implementations. The ORIGINAL paper introduces a block-diagonal mask mechanism for attention layers, while ShotAdapter[3] uses 'a local attention masking strategy which controls the transition token's effect.' Both approaches aim to achieve frame-level control of transitions in multi-shot videos without requiring extensive retraining, indicating that the core concept of using attention masks for transition control was not first proposed by the ORIGINAL paper.

Evidence

Evidence 1 - **Rationale**: Both papers describe using attention masking mechanisms to control shot transitions at specific frame positions. The ORIGINAL paper's 'mask mechanism featuring strong correlations within shots and weak correlations between shots' parallels ShotAdapter[3]'s 'local attention masking strategy which controls the transition token's effect.' - **Original**: we introduce a mask mechanism featuring strong correlations within shots and weak correlations between shots, achieving controlled cinematic transitions and enabling zero-shot multi-shot generation - **Candidate**: This is achieved by incorporating a transition token into the text-to-video model to control at which frames a new shot begins and a local attention masking strategy which controls the transition token's effect and allows shot-specific prompting

Evidence 2 - **Rationale**: Both papers demonstrate that their mask-based approaches can be applied to pre-trained models with minimal additional training. The ORIGINAL paper's 'training-free setting' and ShotAdapter[3]'s 'fine-tuning... for a few thousand iterations' both indicate lightweight adaptation methods for enabling multi-shot generation. - **Original**: we introduce a mask mechanism that enables cinematic transitions within diffusion models, leading to the cinetrans framework, which is effective in a training-free setting - **Candidate**: Extensive experiments demonstrate that fine-tuning a pre-trained text-to-video model for a few thousand iterations is enough for the model to subsequently be able to generate multi-shot videos with shot-specific control

Evidence 3 - **Rationale**: Both papers describe mechanisms for controlling multi-shot video generation through attention manipulation. The ORIGINAL paper's approach of 'strong correlations within shots and weak correlations between shots' and ShotAdapter[3]'s 'full attention across all frames of all shots' with 'shot-specific conditioning' represent similar conceptual approaches to managing intra-shot versus inter-shot relationships. - **Original**: Building on this, we introduce a mask mechanism featuring strong correlations within shots and weak correlations between shots, achieving controlled cinematic transitions and enabling zero-shot multi-shot generation - **Candidate**: our approach enables generation of multi-shot videos as a single video with full attention across all frames of all shots, ensuring character and background consistency, and allows users to control the number, duration, and content of shots through shot-specific conditioning

---

10. Seine: Short-to-long video diffusion model for generative transition and prediction

URL: View paper

Prior Art Analysis

Seine[54] demonstrates prior work on mask-based mechanisms for video diffusion models that control frame-level transitions. Seine[54] proposes a 'random-mask video diffusion model' that uses binary masks to selectively preserve or suppress information from frames, enabling control over which frames are visible during generation. This mechanism operates on attention layers to generate transition videos between scenes, similar to the original paper's block-diagonal mask approach. Both methods apply masks to attention modules in diffusion models to control temporal correlations and achieve frame-level transition control, with Seine[54] published at ICLR 2024, predating the original submission.

Evidence

Evidence 1 - **Rationale**: Both papers propose mask mechanisms in diffusion models for controlling frame-level transitions. Seine[54] introduces a 'random mask module' for transition generation, while the original paper introduces a 'mask mechanism' for cinematic transitions. - **Original**: we introduce a mask mechanism featuring strong correlations within shots and weak correlations between shots, achieving controlled cinematic transitions and enabling zero-shot multi-shot generation - **Candidate**: we develop a diffusion-based model for s2l videos via transition generation and video prediction. in order to generate unseen frames of transition and prediction, based on visible conditional images or videos, our s2l video diffusion model incorporates a random mask module

Evidence 2 - **Rationale**: Seine[54] explicitly describes using binary masks to control which frames are visible/masked in the diffusion process, achieving transition control between scenes, which is functionally similar to the original paper's mask mechanism for shot transitions. - **Original**: Building on this, we introduce a mask mechanism featuring strong correlations within shots and weak correlations between shots, achieving controlled cinematic transitions and enabling zero-shot multi-shot generation. - **Candidate**: we introduce a randommask condition layer at the input stage. this layer applies a binary mask m ∈ rn broadcasting to ˜m ∈ rnxcxhxw as the size of the latent code, resulting in a masked latent code: ˜z0 = z0 ⊙ m, ˜z0 ∈ rnxcxhxw. the binary masks, represented by m, serve as a mechanism to selectively...

Evidence 3 - **Rationale**: Both papers demonstrate that their mask mechanisms enable precise frame-level control over transitions. Seine[54] shows this by generating transitions between arbitrary frame positions, similar to the original paper's frame-level transition control. - **Original**: the application of mask enables strong intra-shot frame correlations in attention module, facilitating precise frame-level cinematic transitions, which remains effective even in a training-free setting - **Candidate**: our randommask based model is capable of generating frames for any given frames at arbitrary positions within the sequence. transition can be obtained by providing the first and last frames of a sequence and utilizing prompts to control the transition style and content

---

Contribution 3: CineTrans framework for multi-shot video generation

Description: The authors propose CineTrans, a framework that combines the mask mechanism with fine-tuning on Cine250K to generate multi-shot videos with cinematic transitions. The framework produces videos that adhere to film editing conventions while maintaining temporal consistency.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. FilMaster: Bridging Cinematic Principles and Generative AI for Automated Film Generation

URL: View paper

Brief Assessment

FilMaster[10] focuses on end-to-end film production with camera language design and post-production workflows, rather than specifically addressing cinematic transitions via masked diffusion models as CineTrans does.

---

2. MovieFactory: Automatic Movie Creation from Text using Large Generative Models for Language and Images

URL: View paper

Brief Assessment

MovieFactory[17] focuses on automated movie creation with multi-scene generation using ChatGPT for script expansion and retrieval-based audio, rather than cinematic transitions with mask mechanisms. The technical approaches differ fundamentally in their control mechanisms and objectives.

---

3. Cine-AI: Generating Video Game Cutscenes in the Style of Human Directors

URL: View paper

Brief Assessment

Cine-AI[53] focuses on procedural cinematography for in-game cutscenes using a timeline/storyboard interface in Unity, not on multi-shot video generation with diffusion models and cinematic transitions as in CineTrans.

---

4. CineLOG: A Training Free Approach for Cinematic Long Video Generation

URL: View paper

Brief Assessment

CineLOG[20] focuses on camera trajectory control and genre-based generation using a decoupled pipeline with trajectory-guided transitions, whereas the original paper addresses cinematic transitions through mask-based diffusion with fine-tuning on Cine250K. The technical approaches and core objectives differ substantially.

---

5. MoCha: Towards Movie-Grade Talking Character Synthesis

URL: View paper

Brief Assessment

MoCha[52] focuses on generating talking character animations from speech and text, emphasizing character-driven storytelling and speech-video synchronization. This is fundamentally different from CineTrans's focus on cinematic transitions and film editing conventions in multi-shot video generation.

---

6. Skyreels-v2: Infinite-length film generative model

URL: View paper

Brief Assessment

Skyreels v2[51] focuses on infinite-length film generation using diffusion forcing and reinforcement learning, not specifically on cinematic transitions via masked diffusion models as in CineTrans.

---

7. ShotDirector: Directorially Controllable Multi-Shot Video Generation with Cinematographic Transitions

URL: View paper

Prior Art Analysis

ShotDirector[15] demonstrates that prior work exists in multi-shot video generation with cinematic transitions. Both papers address the same core problem: generating multi-shot videos with professional film-style transitions using diffusion models. ShotDirector[15] proposes a framework that combines camera control and hierarchical prompting to achieve controllable shot transitions with professional editing patterns (cut-in, cut-out, shot/reverse shot, multi-angle), trained on a dataset with detailed transition annotations. The paper explicitly positions itself as addressing limitations in existing multi-shot generation methods, including those that use mask mechanisms. The temporal overlap and similar technical approach (mask-based control, specialized datasets, hierarchical captions) suggest that the novelty claim of being the first to propose a framework for cinematic multi-shot video generation may be challenged.

Evidence

Evidence 1 - **Rationale**: Both papers propose frameworks for multi-shot video generation with cinematic transitions. ShotDirector[15] explicitly addresses the same problem space with a mask mechanism and hierarchical prompting, suggesting prior work exists in this area. - **Original**: we introduce cinetrans, a novel framework for generating coherent multi-shot videos with cinematic, film-style transitions. to facilitate insights into the film editing style, we construct a multi-shot video-text datasetcine250kwith detailed shot annotations. - **Candidate**: we proposeshotdirector, an efficient framework that integrates parameter-level camera control and hierarchical editing-pattern-aware prompting. specifically, we adopt a camera control module that incorporates 6-dof poses and intrinsic settings to enable precise camera information injection. in addit...

Evidence 2 - **Rationale**: Both papers construct specialized datasets with hierarchical annotations for training multi-shot video generation models with cinematic transitions, indicating similar approaches to the problem. - **Original**: we developed a dataset of 250k video-text pairs, complete with frame-level shot labels and hierarchical annotations, which facilitates video diffusion models for generating cinematic transitions and consistency between shots. - **Candidate**: to train our framework, we construct a high-quality multi-shot video dataset shotweaver40k. a rigorous data curation pipeline is adopted to carefully filter cinematic raw footage, ensuring that transitions between shots follow plausible narrative or spatial reasoning rather than arbitrary visual cha...

Evidence 3 - **Rationale**: Both papers employ mask mechanisms to control shot transitions in diffusion models, with ShotDirector[15] providing a similar technical approach to achieving cinematic multi-shot generation. - **Original**: we analyze attention maps in diffusion models for multi-shot video generation and observe a strong connection between attention probabilities and shot transitions. building on this insight, we introduce a mask mechanism that enables cinematic transitions within diffusion models - **Candidate**: we propose a shot-aware mask mechanism that guides context modeling at both global and local levels, enabling fine-grained control over token interactions. by structuring the visibility of tokens, it allows the model to balance global coherence with shot-specific diversity. this mechanism aligns tex...

---

8. STAGE: Storyboard-Anchored Generation for Cinematic Multi-shot Narrative

URL: View paper

Brief Assessment

STAGE[24] uses a storyboard-anchored approach with start-end frame pairs and focuses on keyframe-based methods, while CineTrans employs a mask mechanism with diffusion models for direct multi-shot generation. These represent fundamentally different technical approaches to multi-shot video generation.

---

9. VideoGen-of-Thought: Step-by-step generating multi-shot video with minimal manual intervention

URL: View paper

Brief Assessment

VideoGen of Thought[1] focuses on narrative-driven multi-shot generation with identity preservation and storyline modeling, while the original paper emphasizes cinematic transitions through mask mechanisms and film-editing conventions. These represent distinct technical approaches to multi-shot video generation.

---

10. Cut2Next: Generating Next Shot via In-Context Tuning

URL: View paper

Brief Assessment

Cut2Next[12] focuses on next shot generation (NSG) with editing pattern adherence, while CineTrans addresses multi-shot generation with cinematic transitions via masked diffusion. The tasks and technical approaches differ fundamentally.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] CineTrans: Learning to Generate Videos with Cinematic Transitions via Masked Diffusion Models View paper
- [1] VideoGen-of-Thought: Step-by-step generating multi-shot video with minimal manual intervention View paper
- [2] CineVerse: Consistent Keyframe Synthesis for Cinematic Scene Composition View paper
- [3] ShotAdapter: Text-to-Multi-Shot Video Generation with Diffusion Models View paper
- [4] Shot sequence ordering for video editing: Benchmarks, metrics, and cinematology-inspired computing methods View paper
- [5] StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text View paper
- [6] MotionCanvas: Cinematic Shot Design with Controllable Image-to-Video Generation View paper
- [7] Can video generation replace cinematographers? Research on the cinematic language of generated video View paper
- [8] Holocine: Holistic generation of cinematic multi-shot long video narratives View paper
- [9] TalkCuts: A Large-Scale Dataset for Multi-Shot Human Speech Video Generation View paper
- [10] FilMaster: Bridging Cinematic Principles and Generative AI for Automated Film Generation View paper
- [11] MAVIN: Multi-Action Video Generation with Diffusion Models via Transition Video Infilling View paper
- [12] Cut2Next: Generating Next Shot via In-Context Tuning View paper
- [13] DiTCtrl: Exploring Attention Control in Multi-Modal Diffusion Transformer for Tuning-Free Multi-Prompt Longer Video Generation View paper
- [14] STORYANCHORS: Generating Consistent Multi-Scene Story Frames for Long-Form Narratives View paper
- [15] ShotDirector: Directorially Controllable Multi-Shot Video Generation with Cinematographic Transitions View paper
- [16] Text2Story: Advancing Video Storytelling with Text Guidance View paper
- [17] MovieFactory: Automatic Movie Creation from Text using Large Generative Models for Language and Images View paper
- [18] Dynamic Storyboard Generation in an Engine-based Virtual Environment for Video Production View paper
- [19] Trans4D: Realistic Geometry-Aware Transition for Compositional Text-to-4D Synthesis View paper
- [20] CineLOG: A Training Free Approach for Cinematic Long Video Generation View paper
- [21] Infinite Video Generation with Cinematic Camera Trajectory Control View paper
- [22] Fine-Tuning Open Video Generators for Cinematic Scene Synthesis: A Small-Data Pipeline with LoRA and Wan2.1 I2V View paper
- [23] EditIQ: Automated Cinematic Editing of Static Wide-Angle Videos via Dialogue Interpretation and Saliency Cues View paper
- [24] STAGE: Storyboard-Anchored Generation for Cinematic Multi-shot Narrative View paper
- [25] MultiShotMaster: A Controllable Multi-Shot Video Generation Framework View paper
- [26] Cinematographic-Aware Coherent Shot Assembly for How-To Vlog Generation View paper
- [27] Multimodal Cinematic Video Synthesis Using Text-to-Image and Audio Generation Models View paper
- [28] Wan-S2V: Audio-Driven Cinematic Video Generation View paper
- [29] TeViS: Translating Text Synopses to Video Storyboards View paper
- [30] Multi-Conditional Generative Adversarial Network for Text-to-Video Synthesis View paper
- [31] LoCoT2V-Bench: A Benchmark for Long-Form and Complex Text-to-Video Generation View paper
- [32] Artificial Intelligence in Multimedia Content Generation: A Review of Audio and Video Synthesis Techniques View paper
- [33] From Loop to Video Essay View paper
- [34] TimeChat-Online: 80% Visual Tokens are Naturally Redundant in Streaming Videos View paper
- [35] MultiCOIN: Multi-Modal COntrollable Video INbetweening View paper
- [36] Fan Video of Attractions: From Loop to Video Essay View paper
- [37] VideoGen-of-Thought: A Collaborative Framework for Multi-Shot Video Generation View paper
- [38] JAWS: Just A Wild Shot for Cinematic Transfer in Neural Radiance Fields View paper
- [39] GAZED- Gaze-guided Cinematic Editing of Wide-Angle Monocular Video Recordings View paper
- [40] JenBridge: Adaptive Long-Form Video Soundtracking across Scene Transition View paper
- [41] SpatialVID: A Large-Scale Video Dataset with Spatial Annotations View paper
- [42] Procedure-Aware Surgical Video-language Pretraining with Hierarchical Knowledge Augmentation View paper
- [43] HecVL: Hierarchical Video-Language Pretraining for Zero-shot Surgical Phase Recognition View paper
- [44] AnimeShooter: A Multi-Shot Animation Dataset for Reference-Guided Video Generation View paper
- [45] LVD-2M: A Long-take Video Dataset with Temporally Dense Captions View paper
- [46] Video recap: Recursive captioning of hour-long videos View paper
- [47] Hierarchical Video-Moment Retrieval and Step-Captioning View paper
- [48] Video paragraph captioning using hierarchical recurrent neural networks View paper
- [49] MovieBench: A Hierarchical Movie Level Dataset for Long Video Generation View paper
- [50] FineBio: A Fine-Grained Video Dataset of Biological Experiments with Hierarchical Annotation View paper
- [51] Skyreels-v2: Infinite-length film generative model View paper
- [52] MoCha: Towards Movie-Grade Talking Character Synthesis View paper
- [53] Cine-AI: Generating Video Game Cutscenes in the Style of Human Directors View paper
- [54] Seine: Short-to-long video diffusion model for generative transition and prediction View paper
- [55] Dreamix: Video diffusion models are general video editors View paper
- [56] FreeTraj: Tuning-Free Trajectory Control in Video Diffusion Models View paper
- [57] Mask^ 2DiT: Dual Mask-based Diffusion Transformer for Multi-Scene Long Video Generation View paper
- [58] Peekaboo: Interactive video generation via masked-diffusion View paper
- [59] Diffusion action segmentation View paper
- [60] TRIP: Temporal Residual Learning with Image Noise Prior for Image-to-Video Diffusion Models View paper
- [61] EIDT-V: Exploiting Intersections in Diffusion Trajectories for Model-Agnostic, Zero-Shot, Training-Free Text-to-Video Generation View paper