Novelty Assessment Report

Paper: Pusa V1.0: Unlocking Temporal Control in Pretrained Video Diffusion Models via Vectorized Timestep Adaptation
PDF URL: https://openreview.net/pdf?id=4adY8FepXg
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

The rapid advancement of video diffusion models has been hindered by fundamental limitations in temporal modeling, particularly the rigid synchronization of frame evolution imposed by conventional scalar timestep variables. While task-specific adaptations and autoregressive models have sought to address these challenges, they remain constrained by computational inefficiency, catastrophic forgetting, or narrow applicability. In this work, we present \textbf{Pusa} V1.0, a versatile model that leverages \textbf{vectorized timestep adaptation (VTA)} to enable fine-grained temporal control within a unified video diffusion framework. Note that VTA is a non-destructive adaptation, which means that it fully preserves the capabilities of the base model. \textbf{Unlike conventional methods like Wan-I2V, which finetune a base text-to-video (T2V) model with abundant resources to do image-to-video (I2V), we achieve comparable results in a zero-shot manner after an ultra-efficient finetuning process based on VTA. Moreover, this method also unlocks many other zero-shot capabilities simultaneously, such as start-end frames and video extension ---all without task-specific training. Meanwhile, it keeps the T2V capability from the base model.} Mechanistic analyses also reveal that our approach preserves the foundation model's generative priors while surgically injecting temporal dynamics, avoiding the combinatorial explosion inherent to the vectorized timestep. This work establishes a scalable, efficient, and versatile paradigm for next-generation video synthesis, democratizing high-fidelity video generation for research and industry alike.

Core Task Landscape

This paper addresses: **Fine-Grained Temporal Control in Video Diffusion Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **24 categories**.

Taxonomy Overview

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

- **Temporal Modeling Architectures and Mechanisms**
- **Conditional Control Mechanisms and Interfaces**
- **Motion and Appearance Customization**
- **Temporal Consistency and Coherence Enhancement**
- **Training and Optimization Strategies**
- **Specialized Video Generation Tasks**
- **Autoregressive and Sequential Generation**
- **Foundational Video Diffusion Models**

Complete Taxonomy Tree

- Fine-Grained Temporal Control in Video Diffusion Models Survey Taxonomy
- Temporal Modeling Architectures and Mechanisms
  - Vectorized and Frame-Level Timestep Control ★ (2 papers)
  - [0] Pusa V1.0: Unlocking Temporal Control in Pretrained Video Diffusion Models via Vectorized Timestep Adaptation (Anon et al., 2026) View paper
  - [19] Redefining temporal modeling in video diffusion: The vectorized timestep approach (Liu Yaofang, 2024) View paper
  - Temporal Attention and Recurrence Mechanisms (3 papers)
  - [16] VMC: Video Motion Customization Using Temporal Attention Adaption for Text-to-Video Diffusion Models (Hyeonho Jeong, 2024) View paper
  - [33] Streaming Video Diffusion: Online Video Editing with Diffusion Models (Chen Feng, 2024) View paper
  - [39] Cross-frame representation alignment for fine-tuning video diffusion models (Hwang Sung-Won, 2025) View paper
  - Space-Time Joint Generation Architectures (2 papers)
  - [7] Lumiere: A space-time diffusion model for video generation (Bar-Tal, 2024) View paper
  - [30] Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models (Andreas Blattmann, 2023) View paper
  - Latent Space Temporal Modeling (2 papers)
  - [26] Learning Spatial Adaptation and Temporal Coherence in Diffusion Models for Video Super-Resolution (Zhikai Chen, 2024) View paper
  - [36] MagicVideo: Efficient Video Generation With Latent Diffusion Models (Zhou, 2022) View paper
- Conditional Control Mechanisms and Interfaces
  - Multi-Modal Spatial Control Integration (3 papers)
  - [1] Enabling versatile controls for video diffusion models (Zhang Xu, 2025) View paper
  - [2] Ccedit: Creative and controllable video editing via diffusion models (Ruoyu Feng, 2024) View paper
  - [4] Structure and content-guided video synthesis with diffusion models (Zhou Lie, 2023) View paper
  - Trajectory and Motion Field Control (3 papers)

    - [12] TrackDiffusion: Tracklet-Conditioned Video Generation via Diffusion Models (Pengxiang Li, 2025) View paper
    - [17] Motionagent: Fine-grained controllable video generation via motion field agent (Zeng XianFang, 2025) View paper
    - [21] DragNUWA: Fine-grained Control in Video Generation by Integrating Text, Image, and Trajectory (Yin, 2023) View paper
    - 3D-Aware and Camera Control (3 papers)
    - [8] Generative rendering: Controllable 4d-guided video generation with 2d diffusion models (Shengqu Cai, 2024) View paper
    - [14] Controlling space and time with diffusion models (Watson, 2024) View paper
    - [18] Diffusion as shader: 3d-aware video diffusion for versatile video generation control (Zekai Gu, 2025) View paper
    - Scene Graph and Object-Level Control (2 papers)
    - [35] Fine-grained Controllable Video Generation via Object Appearance and Context (Hsin-Ping Huang, 2025) View paper
    - [45] Sg2vid: Scene graphs enable fine-grained control for video synthesis (Yannik Frisch, 2025) View paper
    - Zero-Shot and Training-Free Control Adaptation (2 papers)
    - [6] Spatio-temporal energy-guided diffusion model for zero-shot video synthesis and editing (Ling Yang, 2025) View paper
    - [34] Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model (Lin Han, 2024) View paper
- Motion and Appearance Customization
    - Motion Concept Customization (1 papers)
    - [15] MotionDirector: Motion Customization of Text-to-Video Diffusion Models (Rui Zhao, 2023) View paper
    - Motion Transfer and Attention-Based Manipulation (2 papers)
    - [23] MotionFlow: Attention-Driven Motion Transfer in Video Diffusion Models (Tuna Han Salih Meral, 2024) View paper
    - [24] Conditional Image-to-Video Generation with Latent Flow Diffusion Models (Haomiao Ni, 2023) View paper
- Temporal Consistency and Coherence Enhancement
    - Temporal Alignment and Consistency Enforcement (2 papers)
    - [11] Edit temporal-consistent videos with image diffusion model (Yuanzhi Wang, 2024) View paper
    - [28] Upscale-A-Video: Temporal-Consistent Diffusion Model for Real-World Video Super-Resolution (Shangchen Zhou, 2024) View paper
    - Context Propagation and Long-Range Modeling (2 papers)
    - [13] Frame-Level Captions for Long Video Generation with Complex Multi Scenes (Zheng, 2025) View paper
    - [27] Frame Context Packing and Drift Prevention in Next-Frame-Prediction Video Diffusion Models (Zhang Lvmin, 2025) View paper
    - Diffusion Prior-Based Temporal Consistency (2 papers)
    - [31] Learning Temporally Consistent Video Depth from Video Diffusion Priors (Jiahao Shao, 2025) View paper
    - [44] Depthsync: Diffusion guidance-based depth synchronization for scale-and geometry-consistent video depth estimation (Zhao Wang, 2025) View paper
- Training and Optimization Strategies
    - Preference Optimization and Human Feedback (2 papers)
    - [3] DenseDPO: Fine-Grained Temporal Preference Optimization for Video Diffusion Models (Wu, 2025) View paper
    - [42] InstructVideo: Instructing Video Diffusion Models with Human Feedback (Hangjie Yuan, 2024) View paper
    - Efficient Fine-Tuning and Adaptation (2 papers)
    - [9] Temporal In-Context Fine-Tuning for Versatile Control of Video Diffusion Models (Kim Kinam, 2025) View paper
    - [48] Frame-wise Conditioning Adaptation for Fine-Tuning Diffusion Models in Text-to-Video Prediction (Liu Zheyuan, 2025) View paper
    - Sampling and Inference Guidance (1 papers)
    - [32] Spatiotemporal Skip Guidance for Enhanced Video Diffusion Sampling (Junha Hyung, 2025) View paper
- Specialized Video Generation Tasks
    - Video Editing and Manipulation (2 papers)
    - [46] Bivdiff: A training-free framework for general-purpose video synthesis via bridging image and video diffusion models (Fengyuan Shi, 2024) View paper
    - [47] Videodirector: Precise video editing via text-to-video models (Yukun Wang, 2025) View paper
    - Video Interpolation and Extension (2 papers)
    - [10] Tvg: A training-free transition video generation method with diffusion models (Rui Zhang, 2025) View paper
    - [40] Video Interpolation with Diffusion Models (Siddhant Jain, 2024) View paper
    - Conditional Video Prediction and Synthesis (3 papers)
    - [25] MCVD: Masked Conditional Video Diffusion for Prediction, Generation, and Interpolation (Voleti, 2022) View paper
    - [37] Mind the Time: Temporally-Controlled Multi-Event Video Generation (Ziyi Wu, 2025) View paper
    - [38] Easycontrol: Transfer controlnet to video diffusion for controllable generation and interpolation (Wang Cong, 2024) View paper
    - Video Super-Resolution with Temporal Coherence (1 papers)
    - [49] TempDiff: Enhancing Temporalâawareness in Latent Diffusion for Realâ□World Video Superâ□Resolution (Qin Jiang, 2024) View paper
    - Specialized Domain Applications (4 papers)
    - [20] Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models (Hanwen Liang, 2024) View paper
    - [29] Finemogen: Fine-grained spatio-temporal motion generation and editing (Zhang Mingyuan, 2023) View paper
    - [41] VLOGGER: Multimodal Diffusion for Embodied Avatar Synthesis (Enric Corona, 2025) View paper
    - [43] Timeline and boundary guided diffusion network for video shadow detection (Zhou Hai-peng, 2024) View paper
- Autoregressive and Sequential Generation (1 papers)
    - [5] Art-v: Auto-regressive text-to-video generation with diffusion models (Wen-ming, 2024) View paper
- Foundational Video Diffusion Models (2 papers)
    - [22] Video Diffusion Models (Jonathan Ho, 2022) View paper
    - [50] Imagen Video: High Definition Video Generation with Diffusion Models (Ho, 2022) View paper

Narrative

Core task: fine-grained temporal control in video diffusion models. The field has evolved around several complementary directions that together address how to generate coherent, controllable video sequences. Foundational Video Diffusion Models such as Lumiere[7] and Imagen Video[50] establish baseline architectures for temporal generation, while Temporal Modeling Architectures and Mechanisms explore how to represent and manipulate time at different granularities—ranging from frame-level timestep control (e.g., Vectorized

Timestep[19]) to autoregressive and sequential generation strategies that build videos incrementally. Conditional Control Mechanisms and Interfaces introduce diverse input modalities, including text, sketches, and spatial signals, enabling users to steer content more precisely. Motion and Appearance Customization branches focus on disentangling and personalizing dynamic attributes, with works like MotionDirector[15] and MotionFlow[23] targeting motion-specific tuning. Meanwhile, Temporal Consistency and Coherence Enhancement addresses the challenge of maintaining stable object identity and smooth transitions across frames, and Training and Optimization Strategies investigate efficient learning paradigms, including methods like DenseDPO[3] that refine models via preference-based feedback.

A particularly active line of work centers on achieving fine-grained control over when and how motion unfolds. Pusa[0] sits within the Vectorized and Frame-Level Timestep Control cluster, emphasizing per-frame manipulation of diffusion timesteps to modulate temporal dynamics precisely. This approach contrasts with methods that rely on global conditioning or coarse temporal segmentation, such as those in Conditional Control Mechanisms that apply uniform guidance across the entire sequence. Nearby works like Vectorized Timestep[19] share a similar philosophy of frame-wise parameterization, while others in Motion and Appearance Customization (e.g., MotionDirector[15]) focus more on learning reusable motion priors rather than explicit timestep modulation. The interplay between these branches highlights an ongoing tension: whether to embed temporal control directly into the diffusion schedule or to encode it through learned representations and conditioning signals. Pusa[0] exemplifies the former strategy, offering a complementary perspective to appearance-driven and motion-prior methods.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Redefining temporal modeling in video diffusion: The vectorized timestep approach

Authors: Liu Yaofang, Ren, Yumeng, Yaofang Liu, Cun, et al. (22 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Diffusion models have revolutionized image generation, and their extension to video generation has shown promise. However, current video diffusion models~(VDMs) rely on a scalar timestep variable applied at the clip level, which limits their ability to model complex temporal dependencies needed for various tasks like image-to-video generation. To address this limitation, we propose a frame-aware video diffusion model~(FVDM), which introduces a novel vectorized timestep variable~(VTV). Unlike con...

Relationship Analysis

Both papers belong to the Vectorized and Frame-Level Timestep Control category, utilizing independent per-frame noise schedules for temporal control in video diffusion models. They share the core approach of replacing scalar timesteps with vectorized timestep variables to enable fine-grained temporal modeling, and both demonstrate applications in image-to-video generation, video interpolation, and extension. The key difference is that Pusa V1.0 focuses on efficient adaptation of pretrained models through a non-destructive VTA strategy requiring minimal data (4K samples) and compute (0.5K cost), while FVDM (the candidate paper) introduces the foundational framework with probabilistic timestep sampling strategy (PTSS) and trains models from scratch with comprehensive ablation studies on sampling probabilities and model scales.

Contributions Analysis

Overall novelty summary. The paper introduces Vectorized Timestep Adaptation (VTA) to enable fine-grained temporal control in video diffusion models, positioning itself within the 'Vectorized and Frame-Level Timestep Control' leaf of the taxonomy. This leaf contains only two papers total, including the original work, indicating a relatively sparse research direction. The core contribution centers on using independent per-frame noise schedules rather than scalar timesteps, allowing the model to achieve zero-shot image-to-video generation and other tasks without task-specific training. This approach contrasts with the broader field's tendency toward learned motion priors or global conditioning strategies.

The taxonomy reveals that temporal control research has diversified across multiple branches. Neighboring leaves include 'Temporal Attention and Recurrence Mechanisms' (three papers) and 'Space-Time Joint Generation Architectures' (two papers), which address coherence through architectural components rather than timestep manipulation. The 'Conditional Control Mechanisms' branch explores spatial guidance through edges, trajectories, and 3D priors, while 'Motion and Appearance Customization' focuses on learning reusable motion concepts. Pusa's vectorized timestep approach represents a distinct paradigm: embedding temporal control directly into the diffusion schedule rather than through learned representations or external conditioning signals.

Among twenty-eight candidates examined, the contribution-level analysis reveals mixed novelty signals. The VTA mechanism itself (Contribution 1) examined ten candidates with one appearing to provide overlapping prior work, suggesting some precedent for vectorized timestep concepts within the limited search scope. The unified multi-task framework (Contribution 2) examined nine candidates with one refutable match, indicating that zero-shot generalization approaches exist in related contexts. The Frame-Aware Flow Matching objective (Contribution 3) examined nine candidates with no clear refutations, suggesting this training formulation may be more distinctive within the sampled literature.

Based on the limited search scope of twenty-eight semantically similar papers, the work appears to occupy a sparsely populated research direction with some conceptual overlap in vectorized timestep ideas but potentially novel integration and training strategies. The analysis does not cover exhaustive prior work in video diffusion or temporal control more broadly, and the taxonomy structure suggests active parallel development in complementary approaches to temporal modeling.

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

Contribution 1: Vectorized Timestep Adaptation (VTA) for efficient video diffusion model adaptation

Description: The authors propose a non-destructive adaptation method called Vectorized Timestep Adaptation (VTA) that inflates the scalar timestep variable of pretrained video diffusion models into a frame-level vector. This enables fine-grained temporal control while fully preserving the base model's capabilities, achieving state-of-the-art image-to-video performance with minimal training data and computational cost.

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. VMC: Video Motion Customization Using Temporal Attention Adaption for Text-to-Video Diffusion Models

URL: View paper

Brief Assessment

VMC[16] focuses on customizing motion patterns in video generation through temporal attention adaptation, not on vectorized timestep mechanisms for model adaptation. The candidate addresses motion customization from reference videos, while the original contribution concerns frame-level timestep control for temporal tasks like image-to-video generation.

---

2. MotionDirector: Motion Customization of Text-to-Video Diffusion Models

URL: View paper

Brief Assessment

MotionDirector[15] focuses on motion customization using dual-path LoRAs to decouple appearance and motion learning, not on timestep-level temporal control mechanisms for video diffusion models.

---

3. Timestep Embedding Tells: Itâ□□s Time to Cache for Video Diffusion Model

URL: View paper

Brief Assessment

Timestep Embedding Cache[60] focuses on inference acceleration through caching strategies based on timestep embeddings, not on model adaptation or training methods. The candidate addresses computational efficiency during generation, while the original contribution concerns a training-free adaptation framework that modifies how timestep variables control frame evolution.

---

4. Temporal In-Context Fine-Tuning for Versatile Control of Video Diffusion Models

URL: View paper

Brief Assessment

Temporal In-Context[9] focuses on temporal concatenation of condition and target frames with buffer frames for conditional video generation tasks, rather than vectorized timestep adaptation. The methods address different technical challenges in video diffusion model control.

---

5. Fine-tuning Diffusion Policies with Backpropagation Through Diffusion Timesteps

URL: View paper

Brief Assessment

Backpropagation Through Timesteps[62] focuses on fine-tuning diffusion policies for decision-making tasks (robotics, gaming) using noise-conditioned inference and gradient backpropagation through denoising steps. The original paper addresses video generation with frame-level timestep vectors for temporal control. These are fundamentally different application domains and technical approaches.

---

6. Pioneering 4-Bit FP Quantization for Diffusion Models: Mixup-Sign Quantization and Timestep-Aware Fine-Tuning

URL: View paper

Brief Assessment

Mixup-Sign Quantization[63] focuses on 4-bit floating-point quantization techniques for diffusion models, not on temporal control or adaptation methods for video generation. The candidate addresses model compression through quantization, while the original contribution concerns temporal modeling via vectorized timesteps.

---

7. Redefining temporal modeling in video diffusion: The vectorized timestep approach

URL: View paper

Prior Art Analysis

Vectorized Timestep[19] demonstrates that the core concept of using vectorized timesteps for video diffusion models was previously introduced by FVDM (Frame-aware Video Diffusion Model). The candidate paper explicitly presents the vectorized timestep variable (VTV) as a fundamental innovation that enables independent frame evolution in video diffusion models. Both papers utilize the same mathematical formulation where each frame has its own timestep component τ_i, forming a vector $\tau = [\tau_1, \tau_2, ..., \tau_n]$. The candidate paper's work predates the original paper's submission and establishes the foundational paradigm that the original paper claims as novel.

Evidence

Evidence 1 - **Rationale**: Both papers claim to introduce vectorized timesteps as a novel contribution. The candidate paper explicitly states this is a 'novel vectorized timestep variable' that enables independent frame evolution, which is the same core concept claimed by the original paper's VTA. - **Original**: we presentpusa 1 v1.0, a versatile model that leveragesvectorized timestep adaptation (vta)to enable finegrained temporal control within a unified video diffusion framework. note that vta is a non-destructive adaptation, which means that it fully preserves the capabilities of the base model. - **Candidate**: we introduce a novel frame-aware video diffusion model (fvdm), which introduces a novel vectorized timestep variable (vtv). unlike conventional vdms, our approach allows each frame to follow an independent noise schedule, enhancing the model's capacity to capture fine-grained temporal dependencies.

Evidence 2 - **Rationale**: The original paper acknowledges FVDM's prior work on vectorized timesteps. The candidate paper provides the exact mathematical formulation of the vectorized timestep variable that enables independent frame evolution, demonstrating prior art for this concept. - **Original**: concurrently, fvdm liu et al. (2024b) proposed a vectorized timestep to reframe the video diffusion paradigm fundamentally. specifically, instead of using a scalar timestep to control the noise level of the whole video, it allows independent noise evolution per frame by assigning each frame a timest... - **Candidate**: we introduce a vectorized timestep variable τptq: r0, tsñr 0, tsn , defined as: τptq=r τp1qptq, τp2qptq, . . . , τpnqptqsj where n is the number of frames in the video sequence, and τpiqptqrepresents the individual time variable for the i-th frame. this vectorization allows for independent noise per...

Evidence 3 - **Rationale**: The original paper explicitly states it 'extends the paradigm of frame-aware video diffusion models (fvdm)' which already proposed vectorized timesteps. The candidate paper establishes this paradigm as its core innovation, demonstrating that the vectorized timestep approach existed prior to the original paper's claimed novelty. - **Original**: in this work, we extend the paradigm of frame-aware video diffusion models (fvdm)liu et al. (2024b) to an industrial scale by proposing a vectorized timestep adaptation (vta) strategy, which adapts pretrained large-scale vdms to support frame-level timesteps - **Candidate**: at the heart of our approach lies a vectorized timestep variable (vtv) that enables independent frame evolution (shown in fig. 1(c)). this stands in stark contrast to existing vdms, which rely on a scalar timestep variable that enforces uniform temporal dynamics across all frames.

---

8. Tuning Timestep-Distilled Diffusion Model Using Pairwise Sample Optimization

URL: View paper

Brief Assessment

Pairwise Sample Optimization[65] focuses on fine-tuning timestep-distilled diffusion models for image generation tasks using pairwise preference optimization. It does not address video diffusion models or temporal control mechanisms like VTA.

---

9. Adadiff: Adaptive step selection for fast diffusion

URL: View paper

Brief Assessment

Adadiff[64] focuses on adaptive step selection for diffusion models to balance inference speed and quality, not on vectorized timestep adaptation for video diffusion models. The candidate addresses a different problem domain (step count optimization) rather than frame-level temporal control.

---

10. Simda: Simple diffusion adapter for efficient video generation

URL: View paper

Brief Assessment

Simda[61] focuses on parameter-efficient adaptation using spatial/temporal adapters and latent-shift attention for video generation, not on vectorized timestep control mechanisms. The candidate's approach inflates 2D models to 3D and uses adapters, fundamentally different from VTA's frame-level timestep vectors.

Contribution 2: Unified multi-task video generation framework with zero-shot generalization

Description: The authors develop a unified framework that simultaneously supports multiple video generation tasks including text-to-video, image-to-video, start-end frame conditioning, and video extension without requiring task-specific retraining. This zero-shot multi-task capability emerges from the flexible vectorized timestep control mechanism.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators

URL: View paper

Brief Assessment

Text2Video-Zero[70] focuses on zero-shot text-to-video generation by adapting text-to-image models without training, targeting tasks like text-to-video and video editing. The original paper (PUSA) presents a framework for temporal control in pretrained video diffusion models using vectorized timestep adaptation, enabling tasks like image-to-video, start-end frames, and video extension. These represent fundamentally different technical approaches and task scopes.

---

2. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation

URL: View paper

Brief Assessment

Rerender A Video[73] focuses on video-to-video translation using adapted image diffusion models with temporal consistency constraints, not a unified multi-task video generation framework supporting text-to-video, image-to-video, and video extension tasks.

---

3. Univideo: Unified understanding, generation, and editing for videos

URL: View paper

Brief Assessment

Univideo[71] focuses on a dual-stream architecture combining MLLM and mmDiT for unified understanding, generation, and editing across video tasks. While both papers address multi-task video generation, Univideo[71] emphasizes multimodal instruction following and editing capabilities rather than the vectorized timestep control mechanism that enables PUSA's zero-shot generalization.

---

4. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation

URL: View paper

Brief Assessment

StoryDiffusion[74] focuses on maintaining consistent subjects across generated images/videos for visual storytelling, not on unified multi-task video generation frameworks. The candidate addresses subject consistency through self-attention mechanisms rather than temporal control via vectorized timesteps for diverse video generation tasks.

---

5. Univid: Unifying vision tasks with pre-trained video generation models

URL: View paper

Brief Assessment

Univid[72] focuses on adapting pre-trained video generation models to diverse vision tasks (including understanding tasks like segmentation and depth estimation) through visual sentence paradigms. The original paper's contribution centers on temporal control in video diffusion via vectorized timestep adaptation for generation tasks (T2V, I2V, video extension). These represent fundamentally different technical approaches and application domains.

---

6. Fatezero: Fusing attentions for zero-shot text-based video editing

URL: View paper

Brief Assessment

FateZero[69] focuses on zero-shot text-based video editing (style, attribute, shape editing) using pre-trained diffusion models, not on multi-task video generation with tasks like text-to-video, image-to-video, and video extension as described in the original contribution.

---

7. IM-Zero: Instance-level Motion Controllable Video Generation in a Zero-shot Manner

URL: View paper

Brief Assessment

IM-Zero[68] focuses on instance-level motion control through spatial layouts and trajectories for video generation, not on multi-task video generation capabilities like text-to-video, image-to-video, and video extension within a unified framework.

---

8. Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models

URL: View paper

Brief Assessment

Align Your Latents[30] focuses on adapting image LDMs to video generation through temporal alignment layers, primarily for text-to-video and driving simulation tasks. It does not present a unified framework supporting multiple video generation tasks (text-to-video, image-to-video, start-end frame conditioning, video extension) simultaneously with zero-shot generalization as claimed in the original paper's vectorized timestep control mechanism.

---

9. VideoPoet: A Large Language Model for Zero-Shot Video Generation

URL: View paper

Prior Art Analysis

VideoPoet[66] demonstrates a unified multi-task video generation framework that performs multiple tasks (text-to-video, image-to-video, video editing, stylization, etc.) within a single model without task-specific retraining. The paper explicitly describes zero-shot capabilities across diverse tasks through its LLM-based architecture and multi-task pretraining strategy. This directly challenges the novelty claim of

the original paper, as VideoPoet[66] was published prior to the original submission and demonstrates the same core capability of unified multi-task video generation with zero-shot generalization.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly claim zero-shot generalization capabilities. VideoPoet[66] demonstrates that the model can handle new tasks not included in training and process inputs diverging from training distribution, which is the same zero-shot multi-task capability claimed by the original paper. - **Original**: pusa not only preserves t2v capability from the base model, but also generalizes well to many advanced temporal tasks like i2v , start-end frames, video extension, etc., all in a zero-shot way. - **Candidate**: we use the term 'zero-shot video generation' as videopoet processes new text, image, or video inputs that diverge from the training data distribution. furthermore, videopoet handles new tasks not included in its training. For example, videopoet is able to perform new editing tasks by sequentially ch...

Contribution 3: Frame-Aware Flow Matching objective for vectorized timestep training

Description: The authors extend the Frame-Aware Video Diffusion Model paradigm to the flow matching framework by introducing a Frame-Aware Flow Matching (FAFM) objective. This formulation enables each video frame to evolve independently along its own probability path with frame-specific timesteps, avoiding the rigid synchronization of conventional video diffusion models.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Efficient video prediction via sparsely conditioned flow matching

URL: View paper

Brief Assessment

Sparsely Conditioned Flow[59] focuses on video prediction using sparse random frame conditioning during flow matching integration steps, not on frame-aware training with vectorized timesteps for video diffusion models.

---

2. Depth Any Video with Scalable Synthetic Data

URL: View paper

Brief Assessment

Depth Any Video[54] applies flow matching to video depth estimation with spatial-only VAE compression, not to general video diffusion with frame-aware vectorized timesteps. The candidate focuses on depth prediction tasks rather than the temporal control framework proposed in the original paper.

---

3. BAgger: Backwards Aggregation for Mitigating Drift in Autoregressive Video Diffusion Models

URL: View paper

Brief Assessment

BAgger[56] addresses exposure bias in autoregressive video models through backwards aggregation training, not frame-aware flow matching with vectorized timesteps. The candidate focuses on corrective trajectories from model rollouts rather than independent frame-level timestep evolution.

---

4. Conditional Image-to-Video Generation with Latent Flow Diffusion Models

URL: View paper

Brief Assessment

Latent Flow[24] operates in a fundamentally different paradigm - it uses optical flow warping in latent space for image-to-video generation, not frame-aware flow matching with vectorized timesteps. The candidate's approach involves training a flow predictor and diffusion model for motion generation through warping, whereas the original paper proposes independent frame evolution with frame-specific timesteps in a flow matching framework.

---

5. Pyramidal Flow Matching for Efficient Video Generative Modeling

URL: View paper

Brief Assessment

Pyramidal Flow Matching[51] focuses on spatial and temporal pyramid representations for efficient video generation, not on frame-aware flow matching with vectorized timesteps. The candidate uses piecewise flows across resolution pyramids rather than independent per-frame timestep evolution.

---

6. Vsrdiff: Learning inter-frame temporal coherence in diffusion model for video super-resolution

URL: View paper

Brief Assessment

Vsrdiff[55] focuses on video super-resolution using diffusion models with inter-frame temporal coherence modules (IFAG, PRS, FLC), not on frame-aware flow matching objectives or vectorized timestep training for general video generation tasks.

---

7. Genie Envisioner: A Unified World Foundation Platform for Robotic Manipulation

URL: View paper

Brief Assessment

Genie Envisioner[57] focuses on robotic manipulation using video diffusion models for policy learning and simulation, not on frame-aware temporal modeling or vectorized timestep training for general video generation tasks.

---

8. Moalign: Motion-centric representation alignment for video diffusion models

URL: View paper

Brief Assessment

Moalign[52] focuses on motion-centric representation alignment using optical flow supervision to improve physical plausibility in video diffusion models, not on frame-aware flow matching objectives or vectorized timestep training paradigms.

---

9. RoPECraft: Training-Free Motion Transfer with Trajectory-Guided RoPE Optimization on Diffusion Transformers

URL: View paper

Brief Assessment

RoPECraft[53] focuses on training-free motion transfer via RoPE optimization using flow-matching objectives for trajectory alignment during inference, not on training objectives for frame-aware video diffusion models with vectorized timesteps.

Appendix: Text Similarity Detection

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

References

- [0] Pusa V1.0: Unlocking Temporal Control in Pretrained Video Diffusion Models via Vectorized Timestep Adaptation View paper
- [1] Enabling versatile controls for video diffusion models View paper
- [2] Ccedit: Creative and controllable video editing via diffusion models View paper
- [3] DenseDPO: Fine-Grained Temporal Preference Optimization for Video Diffusion Models View paper
- [4] Structure and content-guided video synthesis with diffusion models View paper
- [5] Art-v: Auto-regressive text-to-video generation with diffusion models View paper
- [6] Spatio-temporal energy-guided diffusion model for zero-shot video synthesis and editing View paper
- [7] Lumiere: A space-time diffusion model for video generation View paper
- [8] Generative rendering: Controllable 4d-guided video generation with 2d diffusion models View paper
- [9] Temporal In-Context Fine-Tuning for Versatile Control of Video Diffusion Models View paper
- [10] Tvg: A training-free transition video generation method with diffusion models View paper
- [11] Edit temporal-consistent videos with image diffusion model View paper
- [12] TrackDiffusion: Tracklet-Conditioned Video Generation via Diffusion Models View paper
- [13] Frame-Level Captions for Long Video Generation with Complex Multi Scenes View paper
- [14] Controlling space and time with diffusion models View paper
- [15] MotionDirector: Motion Customization of Text-to-Video Diffusion Models View paper
- [16] VMC: Video Motion Customization Using Temporal Attention Adaption for Text-to-Video Diffusion Models View paper
- [17] Motionagent: Fine-grained controllable video generation via motion field agent View paper
- [18] Diffusion as shader: 3d-aware video diffusion for versatile video generation control View paper
- [19] Redefining temporal modeling in video diffusion: The vectorized timestep approach View paper
- [20] Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models View paper
- [21] DragNUWA: Fine-grained Control in Video Generation by Integrating Text, Image, and Trajectory View paper
- [22] Video Diffusion Models View paper
- [23] MotionFlow: Attention-Driven Motion Transfer in Video Diffusion Models View paper
- [24] Conditional Image-to-Video Generation with Latent Flow Diffusion Models View paper
- [25] MCVD: Masked Conditional Video Diffusion for Prediction, Generation, and Interpolation View paper
- [26] Learning Spatial Adaptation and Temporal Coherence in Diffusion Models for Video Super-Resolution View paper
- [27] Frame Context Packing and Drift Prevention in Next-Frame-Prediction Video Diffusion Models View paper
- [28] Upscale-A-Video: Temporal-Consistent Diffusion Model for Real-World Video Super-Resolution View paper
- [29] Finemogen: Fine-grained spatio-temporal motion generation and editing View paper
- [30] Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models View paper
- [31] Learning Temporally Consistent Video Depth from Video Diffusion Priors View paper
- [32] Spatiotemporal Skip Guidance for Enhanced Video Diffusion Sampling View paper
- [33] Streaming Video Diffusion: Online Video Editing with Diffusion Models View paper
- [34] Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model View paper
- [35] Fine-grained Controllable Video Generation via Object Appearance and Context View paper
- [36] MagicVideo: Efficient Video Generation With Latent Diffusion Models View paper
- [37] Mind the Time: Temporally-Controlled Multi-Event Video Generation View paper
- [38] Easycontrol: Transfer controlnet to video diffusion for controllable generation and interpolation View paper
- [39] Cross-frame representation alignment for fine-tuning video diffusion models View paper
- [40] Video Interpolation with Diffusion Models View paper
- [41] VLOGGER: Multimodal Diffusion for Embodied Avatar Synthesis View paper
- [42] InstructVideo: Instructing Video Diffusion Models with Human Feedback View paper
- [43] Timeline and boundary guided diffusion network for video shadow detection View paper
- [44] Depthsync: Diffusion guidance-based depth synchronization for scale-and geometry-consistent video depth estimation View paper
- [45] Sg2vid: Scene graphs enable fine-grained control for video synthesis View paper
- [46] Bivdiff: A training-free framework for general-purpose video synthesis via bridging image and video diffusion models View paper
- [47] Videodirector: Precise video editing via text-to-video models View paper
- [48] Frame-wise Conditioning Adaptation for Fine-Tuning Diffusion Models in Text-to-Video Prediction View paper
- [49] TempDiff: Enhancing Temporalâ□awareness in Latent Diffusion for Realâ□World Video Superâ□Resolution View paper
- [50] Imagen Video: High Definition Video Generation with Diffusion Models View paper
- [51] Pyramidal Flow Matching for Efficient Video Generative Modeling View paper
- [52] Moalign: Motion-centric representation alignment for video diffusion models View paper
- [53] RoPECraft: Training-Free Motion Transfer with Trajectory-Guided RoPE Optimization on Diffusion Transformers View paper
- [54] Depth Any Video with Scalable Synthetic Data View paper
- [55] Vsrdiff: Learning inter-frame temporal coherence in diffusion model for video super-resolution View paper
- [56] BAgger: Backwards Aggregation for Mitigating Drift in Autoregressive Video Diffusion Models View paper
- [57] Genie Envisioner: A Unified World Foundation Platform for Robotic Manipulation View paper
- [58] FlowLoss: Dynamic Flow-Conditioned Loss Strategy for Video Diffusion Models View paper
- [59] Efficient video prediction via sparsely conditioned flow matching View paper
- [60] Timestep Embedding Tells: Itâ□s Time to Cache for Video Diffusion Model View paper
- [61] Simda: Simple diffusion adapter for efficient video generation View paper
- [62] Fine-tuning Diffusion Policies with Backpropagation Through Diffusion Timesteps View paper
- [63] Pioneering 4-Bit FP Quantization for Diffusion Models: Mixup-Sign Quantization and Timestep-Aware Fine-Tuning View paper
- [64] Adadiff: Adaptive step selection for fast diffusion View paper
- [65] Tuning Timestep-Distilled Diffusion Model Using Pairwise Sample Optimization View paper
- [66] VideoPoet: A Large Language Model for Zero-Shot Video Generation View paper

- [67] Veggie: Instructional editing and reasoning video concepts with grounded generation View paper
- [68] IM-Zero: Instance-level Motion Controllable Video Generation in a Zero-shot Manner View paper
- [69] Fatezero: Fusing attentions for zero-shot text-based video editing View paper
- [70] Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators View paper
- [71] Unividеo: Unified understanding, generation, and editing for videos View paper
- [72] Univid: Unifying vision tasks with pre-trained video generation models View paper
- [73] Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation View paper
- [74] StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation View paper