Efficient Video Generation: A Survey

www.surveyx.cn

Abstract

Efficient video generation is a burgeoning field that seeks to balance high-quality video output with minimal computational resources. This survey paper explores a spectrum of methodologies, including training-free methods that bypass traditional learning processes, training-based approaches leveraging pre-trained models, and diffusion models that refine video frames through probabilistic processes. The paper also examines video synthesis techniques constructing new sequences from existing data and generative models that learn from large datasets to produce realistic content. Key findings highlight the importance of optimizing computational efficiency while maintaining output fidelity, particularly in applications demanding rapid video generation. Innovations such as the Latent Motion Diffusion Model (LaMD) and EfficientDM framework exemplify advancements in handling high-dimensional data and resource constraints. Challenges remain in maintaining temporal coherence and video quality, especially in long sequences and complex scenes. Future research directions include enhancing model architectures for dynamic scene generation, improving synthesis speed, and integrating additional conditioning data to refine video fidelity. The survey underscores the transformative potential of efficient video generation techniques in managing computational resources and meeting the demands of modern applications, paving the way for more sophisticated and accessible video content creation.

1 Introduction

1.1 Importance of Efficient Video Generation

Efficient video generation is crucial in modern applications, significantly impacting computational resource management and the quality of video content. Techniques such as the Infusion method illustrate the potential for drastically reducing computational demands, enabling the production of high-quality videos with minimal resources [1]. This efficiency is particularly essential for generating high-fidelity long videos, given the high-dimensional nature of video data and the limitations of existing generative models [2].

In multimedia applications, methods like MotionCom enhance realism in automatic and motion-aware image composition while minimizing computational requirements [3]. As video-centric platforms advance, the complexity and time-consuming nature of traditional editing approaches further underscore the need for innovative, resource-efficient techniques [4].

Modern applications also grapple with the challenge of producing high-fidelity videos while managing computational costs [5]. The difficulty of generating temporally coherent videos from text prompts without aligned text-video datasets heightens the importance of efficient generation strategies [6].

Moreover, efficient video generation enhances photorealism and captures intricate interactions, which are vital for minimizing resource usage [7]. In dynamic 3D generation, efficient methods are essential for effective resource management [8].

The significance of efficient video generation extends to audio generation, with methods like Make-An-Audio improving usability by generating high-fidelity audio from diverse inputs [9]. The challenges

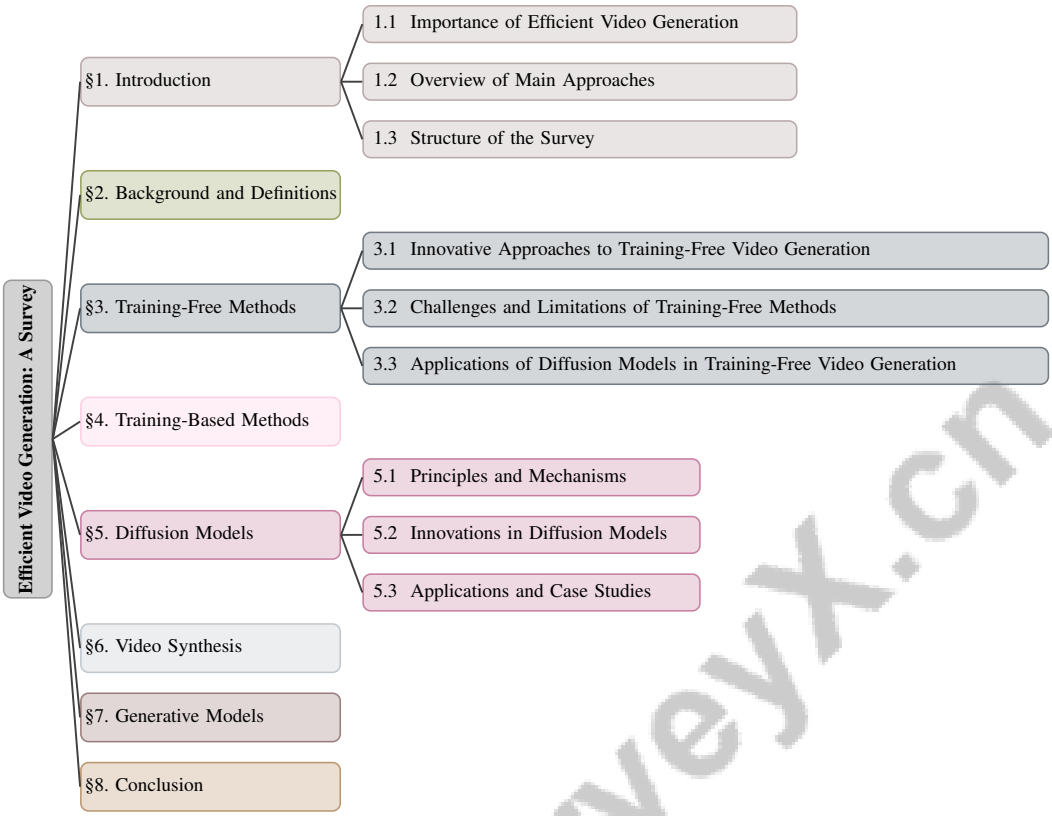


Figure 1: chapter structure

in creating high-quality talking head videos further emphasize the need for efficient generation to manage resources effectively [10].

Diffusion models, known for their computational efficiency in vision tasks, showcase design strategies that balance high-quality outputs with optimized performance [11]. The challenge of generating high-fidelity, temporally coherent videos from text prompts necessitates ongoing advancements in efficient video generation [12]. Collectively, these factors highlight the essential role of efficient video generation in managing computational resources and addressing contemporary application demands.

1.2 Overview of Main Approaches

Efficient video generation encompasses diverse methodologies aimed at optimizing computational resources while maintaining high-quality outputs. Training-free methods, such as the VCUT approach, streamline video generation by eliminating cross-attention mechanisms, thereby reducing computational demands [13]. Complementing these are training-based approaches that leverage pre-trained models, effectively balancing computational costs with video quality through existing knowledge utilization [14].

Diffusion models represent a significant advancement in video generation, with frameworks like Latent Motion Diffusion (LaMD) employing iterative probabilistic processes to refine video frames and enhance motion generation efficiency [15]. The Video Latent Diffusion Model (Video LDM) extends these capabilities by incorporating temporal dimensions to produce long, high-resolution videos [16]. Innovations such as Control-A-Video and Collaborative Video Diffusion (CVD) introduce content and motion priors, along with cross-video synchronization modules, to ensure consistency and coherence in generated videos. Additionally, models like Diffusion4D advance the efficiency of 4D content generation by integrating spatial and temporal consistency [17]. This survey focuses on diffusion models utilized for vision tasks, excluding broader generative models like GANs [11].

Video synthesis techniques address the construction of new video sequences from existing data, tackling temporal coherence and semantic alignment challenges. The Sector-Shaped Diffusion Model (S2DM) exemplifies this by aligning temporal features while preserving semantic characteristics, advancing realistic video content synthesis [18]. ReVideo showcases precise local video editing by integrating content and motion control, highlighting the adaptability of video synthesis methods [19].

Generative models, including diffusion and autoregressive models, are pivotal in video generation, learning patterns from extensive datasets to produce realistic content. These models emphasize the balance between complexity and generative efficiency, as demonstrated by the FreeNoise paradigm, which enhances the capabilities of pre-trained video diffusion models without additional tuning [20]. The Imagen Video system exemplifies text-conditional video generation, utilizing a cascade of video diffusion models to produce high-definition outputs [21].

The field also addresses the computational challenges inherent in traditional video generation methods, which often require significant resources and time. Hybrid video diffusion models, particularly those operating in low-dimensional latent spaces, have shown comparable performance with reduced computational demands [2]. Moreover, integrating object-centric information into video generation, as proposed in the Patch-based Object-centric Video Transformer (POVT), enhances performance and scalability [5].

Recent advancements, such as Make-A-Video, VideoTetris, and ShareGPT4Video, illustrate concerted efforts to enhance efficient video generation. Each method uniquely contributes to the overarching goal of producing high-quality video content while minimizing computational resource usage. For example, Make-A-Video leverages existing text-to-image data to accelerate training and improve video quality, while VideoTetris tackles the complexities of generating intricate scenes through spatio-temporal compositional diffusion. Additionally, ShareGPT4Video enriches video understanding and generation through refined captioning strategies, ultimately enhancing the training data for text-to-video models. These innovations represent significant progress toward achieving dynamic, coherent, and visually appealing video outputs [22, 23, 6, 24].

1.3 Structure of the Survey

The survey is systematically structured to provide a comprehensive examination of efficient video generation methodologies. It begins with an **Introduction** section that emphasizes the significance of efficient video generation in contemporary applications, offering an overview of primary approaches, including training-free methods, training-based methods, diffusion models, video synthesis, and generative models. This section establishes the foundation for the detailed discussions that follow.

The **Background and Definitions** section lays the groundwork by defining key concepts such as video generation and the various methods and models employed in the field, contextualizing the advanced discussions in later sections.

The survey thoroughly investigates , highlighting innovative techniques that eliminate the need for prior learning or adaptation. It emphasizes the strengths and weaknesses of these approaches regarding computational efficiency and output quality. For instance, the Attention-driven Training-free Efficient Diffusion Model (AT-EDM) shows significant efficiency improvements—such as a 38.8

The subsequent section examines **Training-Based Methods**, focusing on the utilization of pre-trained models to enhance video generation efficiency and quality while discussing the trade-offs between computational costs and video quality.

The role of **Diffusion Models** in video generation is thoroughly analyzed, detailing how these models iteratively refine video frames through probabilistic processes and their impact on video quality and computational efficiency.

The survey then addresses , focusing on innovative techniques employed to generate new video sequences from existing data, particularly through video-to-video and text-to-video synthesis. It highlights the challenges of maintaining temporal coherence and visual realism in generated content, as traditional image synthesis approaches often fail to capture motion dynamics, resulting in low-quality outputs. The discussion includes advancements in generative adversarial networks and the use of motion priors from existing video datasets to enhance the realism of synthesized videos,

showcasing significant innovations and methodologies that push the boundaries of video synthesis technology [25, 26].

In the section on **Generative Models**, the paper reviews how these models learn patterns from large datasets to produce realistic video content, emphasizing the balance between model complexity and generation efficiency.

Finally, the **Conclusion** summarizes the key findings of the survey, discusses current trends and challenges, and outlines future research directions in efficient video generation. This structured approach facilitates a comprehensive investigation of the topic, yielding significant insights into cutting-edge diffusion models and their applications in text-to-video generation, particularly in addressing the complexities of compositional prompts and dynamic scene generation. By employing enhanced video data preprocessing and innovative attention mechanisms, this method not only improves accuracy in generated scenes but also bridges the gap between historical data representation and contemporary generative techniques, ultimately enriching the understanding of audiovisual content creation [27, 22].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Video Generation and Its Challenges

Video generation involves crafting dynamic visual sequences that require maintaining both temporal and spatial coherence, demanding sophisticated modeling of high-dimensional data with inherent redundancies. Ensuring temporal consistency is critical for smooth motion transitions and overall video integrity [2]. The challenge intensifies with longer sequences, leading to quality degradation [2]. Integrating static foregrounds with dynamic backgrounds while preserving realism presents additional difficulties [3].

Generating coherent videos from textual descriptions without aligned datasets is particularly challenging, often resulting in poor fidelity and discontinuous motion. This complexity is compounded by the need to model scenes with dynamic backgrounds and multiple moving objects, where existing methods struggle with computational inefficiency and capturing complex interactions [5]. Although diffusion models show promise, their high computational demands limit practical applications [11]. The reliance on intensive mechanisms like Cross-Attention complicates the process further [13]. Achieving precise motion in dynamic 3D models or videos remains a formidable challenge [8].

The task of video deblurring, aimed at recovering high-frequency information from blurry frames, often results in unsatisfactory outputs due to ineffective recovery processes [28]. Generating dynamic 4D scenes frequently lacks photorealism and fails to capture complex interactions due to reliance on limited datasets [7]. Evaluating temporal dynamics in text-to-video models, particularly their alignment with text prompts, remains a critical challenge [23].

In related fields, generating high-fidelity audio from text faces hurdles like scarce quality text-audio pairs and the complexity of modeling long waveforms [9]. Producing high-quality talking head videos synchronized with audio, while generalizing across identities without fine-tuning, illustrates broader video generation difficulties [10]. Generating geometrically consistent novel views from minimal input, such as a single image, while managing ambiguity and extrapolation, presents further challenges [29].

These challenges highlight the complexity of video generation, as developers strive to balance computational efficiency, temporal coherence, and output quality. Advanced evaluation protocols like DEVIL emphasize the need for models to accurately reflect dynamic content aligned with text prompts. Frameworks such as Make-A-Video and VideoTetris exemplify efforts to enhance model capabilities in managing intricate scenes, while large-scale pretrained models like CogVideo address computational and data limitations in text-to-video generation [22, 30, 23, 6]. Continuous research aims to overcome these challenges, enhancing video generation fidelity and efficiency.

2.2 Role of AI in Video Generation

Artificial intelligence (AI) has revolutionized video generation, introducing advanced techniques that enhance both quality and efficiency. The integration of diffusion models, known for success in image generation, marks a pivotal advancement in video tasks. These models iteratively refine frames

through probabilistic processes, addressing temporal consistency and high-quality output challenges. Their application underscores their versatility, bridging generative and discriminative processes for tasks requiring distinct decision boundaries [31].

AI's influence extends to evaluating and enhancing video content, with benchmarks focusing on dynamics, highlighting AI's role in refining generation techniques [23]. AI-driven methods have improved video stability and quality through systematic data curation and structured training, addressing excessive computational overhead [11]. Incorporating AI involves integrating diffusion models with tasks requiring distinct decision boundaries, crucial for developing models that efficiently handle complex video generation without sacrificing quality [31]. Despite advancements, computational inefficiency during training and inference stages remains a challenge, limiting the broader application of diffusion models [11].

In recent years, the exploration of training-free methods in video generation has gained significant traction within the academic community. These methods offer innovative alternatives to traditional training paradigms, enabling researchers to tackle various challenges associated with video creation. Figure 2 illustrates the hierarchical structure of these training-free methods, effectively categorizing innovative approaches, challenges, and applications of diffusion models. Specifically, the figure highlights key methods such as VCUT and ModelScopeT2V, while also addressing challenges including motion estimation and contextual integration. Furthermore, it underscores the pivotal role of diffusion models in enhancing both the quality and efficiency of video generation, thereby providing a comprehensive overview of the current landscape in this rapidly evolving field.

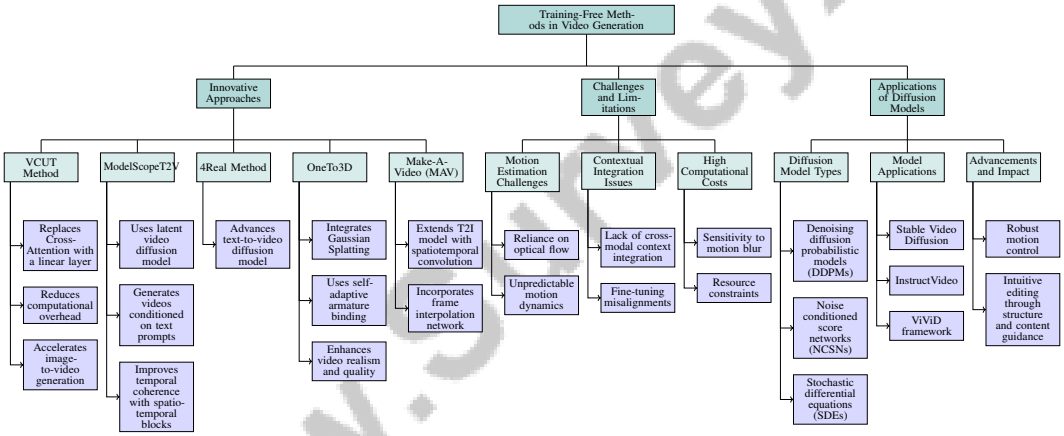


Figure 2: This figure illustrates the hierarchical structure of training-free methods in video generation, categorizing innovative approaches, challenges, and applications of diffusion models. It highlights key methods like VCUT and ModelScopeT2V, addresses challenges such as motion estimation and contextual integration, and explores the pivotal role of diffusion models in enhancing video generation quality and efficiency.

3 Training-Free Methods

3.1 Innovative Approaches to Training-Free Video Generation

Recent developments in training-free video generation emphasize computational efficiency and adaptability. The VCUT method exemplifies this by replacing the Cross-Attention mechanism with a linear layer to reduce computational overhead and accelerate image-to-video generation [13]. This innovation enhances accessibility and efficiency in video creation. ModelScopeT2V employs a latent video diffusion model, generating videos conditioned on text prompts and using spatio-temporal blocks to improve temporal coherence [12]. Similarly, the 4Real method advances training-free video generation through a text-to-video diffusion model [7].

OneTo3D integrates Gaussian Splatting with a self-adaptive armature binding mechanism and re-editable motion control, showcasing adaptive strategies to enhance video realism and quality [8]. Make-A-Video (MAV) extends a T2I model with spatiotemporal convolution and attention layers,

incorporating a frame interpolation network for high frame rate generation [6]. A new evaluation protocol introduces dynamic metrics across multiple temporal granularities, significantly improving the evaluation landscape [23]. These innovative approaches highlight the potential of training-free methods in video generation, offering efficient and flexible solutions that overcome traditional model limitations.

3.2 Challenges and Limitations of Training-Free Methods

Training-free video generation methods, while reducing computational demands, face challenges that limit their efficacy. A primary issue is the reliance on optical flow for motion estimation, which often fails in complex motion scenarios, leading to suboptimal inpainting [1]. Unpredictable motion dynamics in methods like MotionCom can cause inconsistencies due to dependence on multiple seeds for video diffusion models [3]. These methods also struggle with managing multiple objects and interactions, leading to unusual transformations and inaccuracies [22]. The lack of cross-modal context integration further impedes the diffusion process [32], and fine-tuning techniques often produce blurry outputs due to misalignment with the distilled model's capabilities [33].

A significant limitation is disentangling user-provided content edits from structural representations, complicating effective video editing [4]. High computational costs and sensitivity to motion blur exacerbate these issues, resulting in distorted content in deblurred videos [28]. Implicit and explicit methods often lack precise control over dynamic motions due to resource constraints [8]. Inadequate benchmarks misrepresent model capabilities by allowing low-dynamic content generation to achieve high scores [23]. The Make-A-Video method highlights the challenge of associating text with complex phenomena inferred through video [6]. Generative diffusion models, like Make-An-Audio, require substantial computational resources, which may not yield optimal performance with limited data [9].

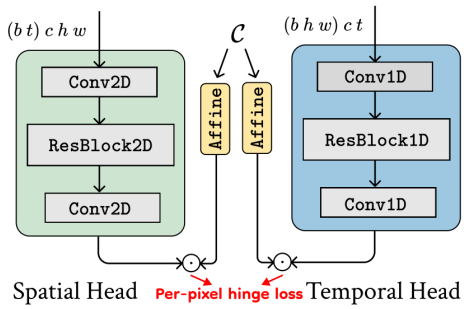
3.3 Applications of Diffusion Models in Training-Free Video Generation

Diffusion models are pivotal in training-free video generation, offering robust solutions for managing high-dimensional sequential data and spatial-temporal correlations. Models such as denoising diffusion probabilistic models (DDPMs), noise conditioned score networks (NCSNs), and stochastic differential equations (SDEs) provide a comprehensive framework for generating videos where traditional training-based approaches are impractical [34]. The Stable Video Diffusion model demonstrates the efficacy of latent diffusion models in producing high-resolution videos from text or image inputs, employing a structured training process to enhance video quality [35]. InstructVideo exemplifies diffusion models' adaptability in refining video outputs based on user input, enhancing text-to-video generation efficiency [36].

The ViViD framework leverages diffusion models to improve video output quality and temporal consistency, particularly in virtual try-on applications [37]. StableMoFusion focuses on efficient high-quality human motion generation, addressing challenges like foot skating through a Conv1D UNet motion-denoising network [38]. EfficientDM optimizes diffusion models for low-bit quantization, advantageous in resource-constrained environments requiring high-quality video generation without additional training data [39]. The LVDM framework exemplifies diffusion models' efficiency in generating longer videos without sacrificing quality [2].

Integrating 3D geometry priors in diffusion-based frameworks extends model capabilities [29], and their application in channel estimation provides a stable training procedure and improved generalization over GAN-based approaches [40]. Advancements in diffusion models underscore their impact on training-free methodologies, enabling intuitive editing through structure and content guidance, facilitating long video generation by breaking down tasks into manageable subtasks, and providing robust motion control without extensive retraining [41, 42, 4].

As depicted in Figure 3, diffusion models are pivotal in training-free video generation, offering significant advancements in producing high-quality content without extensive training. The "Spatial-Temporal Head for 2D Convolutional Networks" enhances video generation by integrating spatial and temporal processing through convolutional networks, optimizing video data dimensions. The "Comparison of User Preferences with and without Image Pretraining" evaluates user preferences across metrics like Prompt Alignment and Quality, highlighting video generation enhancements achieved through pretraining. These examples underscore diffusion models' potential to revolutionize



(a) Spatial-Temporal Head for 2D Convolutional Networks[43]

(b) Comparison of User Preferences with and without Image Pretraining for Prompt Alignment, Quality, and Aggregated Metrics[35]

Figure 3: Examples of Applications of Diffusion Models in Training-Free Video Generation

training-free video generation by leveraging advanced architectures and pretraining techniques [43, 35].

4 Training-Based Methods

4.1 Leveraging Pre-trained Models

Pre-trained models play a pivotal role in video generation by enhancing efficiency and quality through extensive dataset knowledge. Techniques like Make-A-Video exemplify this by utilizing existing text-to-image (T2I) models and unsupervised video learning to produce coherent high-quality videos without needing aligned datasets [6]. This approach effectively streamlines the video generation process.

In vision tasks, efficient diffusion models are categorized into Efficient Design Strategies (EDS) and Efficient Process Strategies (EPS), focusing on architectural modifications and sampling process optimizations, respectively [11]. This classification highlights the role of pre-trained models in refining both structural and procedural aspects, enhancing computational efficiency and output fidelity.

The 4Real method leverages video generative models to improve dynamic scene generation efficiency and quality, contrasting with traditional approaches reliant on fine-tuning pre-trained models on synthetic datasets [7]. This underscores the potential of generative models to surpass conventional pre-trained models' limitations.

Make-An-Audio employs a spectrogram autoencoder and contrastive language-audio pretraining, utilizing large volumes of unsupervised data for audio generation [9]. This illustrates the versatility of pre-trained models in multi-modal contexts, enhancing coherence in generated content.

These approaches demonstrate the strategic use of pre-trained models in video generation, leveraging existing knowledge to ensure high-quality outputs through sophisticated modeling techniques. By using video datasets as motion priors, pre-trained models enhance motion dynamics realism and facilitate efficient video production, addressing common limitations in text-to-video synthesis [22, 25, 6, 26].

4.2 Trade-offs Between Computational Cost and Video Quality

Balancing computational cost and video quality is crucial in training-based video generation, requiring resource optimization alongside output fidelity. Techniques like VCUT reduce latency by up to 20%, optimizing computational expenses while maintaining high-quality video generation [13].

Structured Weight Generation (SWG) addresses limitations of guidance methods like Classifier-Free Guidance (CFG), which often increase training costs and reduce sample diversity. SWG offers a more efficient solution, preserving video quality while managing computational resources [44].

In diffusion models, proposed strategies have significantly improved training efficiency and reduced computational demands, enhancing image generation quality compared to traditional methods [45]. These advancements demonstrate the potential to optimize computational costs without sacrificing content quality.

The ST-Adapter achieves a balance between computational efficiency and performance, matching or exceeding full fine-tuning strategies while significantly reducing parameter counts [46]. However, StableMoFusion illustrates challenges in achieving efficient motion generation, as its inference speed does not yet meet real-time standards [38].

Fourier Diffusion Models outperform scalar diffusion models across various image quality metrics, achieving similar results with fewer time steps, highlighting the trade-off between computational efficiency and video quality [47]. High computational costs associated with training and sampling remain a challenge, particularly for long video generation [35].

Aligning video generation models with human preferences often requires extensive generation from textual inputs, posing significant computational challenges [36]. CustomCrafter addresses this by preserving video diffusion models' inherent capabilities for motion generation and conceptual combination without requiring additional video for fine-tuning, balancing computational cost and video quality [48].

The instability and resource-intensive nature of GAN-based methods present challenges in maintaining image quality and generalization across identities [10]. These challenges highlight ongoing efforts to optimize trade-offs between computational expenses and video quality in training-based methods, striving for efficient and high-quality video generation. The ability to generate diverse and high-quality novel views while maintaining geometrical consistency across frames, as demonstrated by certain methods, showcases the potential to outperform existing regression-based approaches [29].

4.3 Innovative Architectures and Techniques

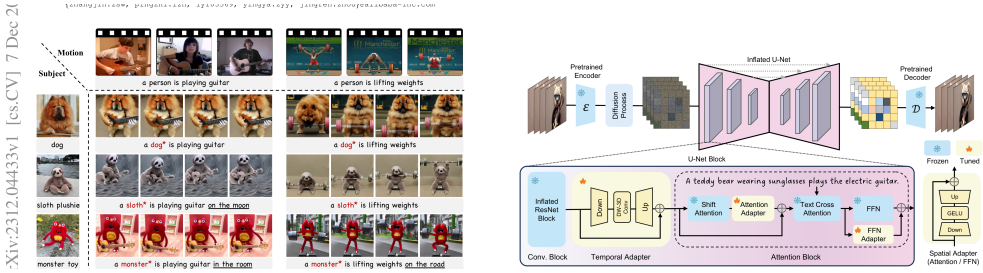
Recent innovations in architecture and techniques have significantly advanced training-based video generation, improving output quality and generation efficiency. StyleGAN-V exemplifies these advancements by integrating continuous motion representations with a novel discriminator design, enhancing video quality and generation efficiency [49].

Hybrid architectures combining diffusion models with various generative techniques have further advanced video generation, enabling intuitive and efficient editing workflows. Innovations like structure and content-guided video diffusion models allow for precise video editing based on textual descriptions, overcoming previous limitations in temporal consistency and structural fidelity. Frameworks like VideoTetris enhance text-to-video generation using spatio-temporal compositional techniques, adeptly managing complex scenes with multiple objects and dynamic changes, thus improving overall quality and coherence [22, 4]. These architectures capitalize on the strengths of each model type, optimizing trade-offs between computational efficiency and output quality.

Adaptive learning strategies and modular architectures in video generation frameworks, such as VideoTetris and CogVideo, have significantly improved models' ability to manage diverse video generation tasks effectively. These advancements enable models to generalize across varied input conditions and complex scenarios, such as dynamic object changes and intricate scene compositions, without extensive retraining. Techniques like spatio-temporal compositional diffusion and multi-frame-rate hierarchical training enhance models' understanding of motion dynamics and textual semantics, leading to more coherent and contextually accurate video outputs [22, 30, 6, 48, 24]. These innovations have been pivotal in reducing the computational burden while maintaining high-quality outputs.

Collectively, these innovative architectures and techniques underscore ongoing efforts to optimize training-based video generation methods, paving the way for efficient and high-quality video content creation. The exploration and implementation of novel designs and strategies are crucial for addressing the intricate challenges of video generation, particularly in enhancing the capabilities of generative models such as Stable Video Diffusion (SVD) and VideoTetris. These advancements aim to improve visual consistency, natural motion, and compositional accuracy in generated videos, ultimately fostering creativity and efficiency in fields like animation, advertising, and educational content creation. By leveraging sophisticated frameworks and evaluation protocols that emphasize dy-

namics, researchers can significantly elevate the quality and applicability of text-to-video generation technologies [22, 23, 50, 6].



(a) Motion and Subject Detection in Video Sequences[51]

(b) Inflated U-Net with Spatial Attention for Text-to-Image Translation[52]

Figure 4: Examples of Innovative Architectures and Techniques

As shown in Figure 4, the exploration of training-based methods and innovative architectures in computer vision and machine learning is exemplified through two intriguing examples. The first, "Motion and Subject Detection in Video Sequences," demonstrates advanced detection techniques in video analysis, showcasing a comparative study where original frames are juxtaposed with detection results, highlighting the system's ability to accurately identify and label subjects interacting with objects. This underscores the sophistication of motion and subject detection algorithms in dynamic environments. The second example, "Inflated U-Net with Spatial Attention for Text-to-Image Translation," presents a novel neural network architecture designed for translating textual descriptions into visual representations. This architecture employs an encoder-decoder structure enhanced with spatial attention mechanisms, enabling effective capturing and reconstruction of intricate details from text inputs into coherent images. Together, these examples illustrate the potential of innovative architectures in enhancing machine learning capabilities and highlight the ongoing evolution of techniques aimed at bridging the gap between textual and visual data representations [51, 52].

5 Diffusion Models

The exploration of diffusion models in the context of video generation reveals a rich tapestry of principles and mechanisms that underpin their functionality. As we delve into the intricacies of these models, it becomes essential to understand the foundational concepts that drive their innovative capabilities. This section will elucidate the core principles and mechanisms that define diffusion models, highlighting their significance in achieving high-quality video generation.

5.1 Principles and Mechanisms

Diffusion models have emerged as a pivotal framework in video generation, employing iterative refinement and probabilistic processes to produce high-quality and temporally coherent video sequences. These models leverage the inherent spatio-temporal correlations among video frames, which are vital for capturing dynamic interactions and ensuring coherence in generated content [7]. The Latent Video Diffusion Model (LVDM) exemplifies this capability by utilizing a low-dimensional latent space to optimize computational efficiency while employing hierarchical modeling and noise conditioning to maintain video quality [2].

The effectiveness of diffusion models is further illustrated by the Infusion method, which capitalizes on the auto-similarity of video content, learning directly from available data without reliance on external datasets [1]. This approach underscores the potential of diffusion frameworks to enhance video quality through intrinsic data properties, thereby reducing dependency on extensive training datasets.

Advanced models such as CONTEXTDIFF introduce innovative mechanisms by contextualizing both forward and reverse diffusion processes through cross-modal interactions, thereby enhancing the integration of multimodal data in video generation [32]. This innovation highlights the adaptability of diffusion models in handling complex video generation tasks that require nuanced data interactions.

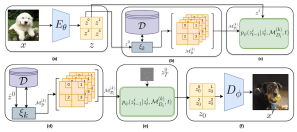
Attention mechanisms play a crucial role in these models, as demonstrated by methodologies like MotionCom, which naturally incorporate objects into video sequences by leveraging learned motion dynamics, ensuring that inserted objects interact seamlessly with the background [3]. This approach exemplifies the potential of attention-based models to refine video content generation.

Moreover, the EfficientDM framework showcases the capability of diffusion models to achieve quantization-aware training (QAT) level performance with post-training quantization (PTQ) level efficiency, underscoring their robustness in low-bit quantization scenarios [39]. This efficiency is essential for maintaining high-quality outputs in resource-constrained environments.

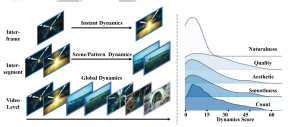
The principles behind 4Real further illustrate the ability of diffusion models to capture dynamic interactions through learned deformations, enhancing realism and structural integrity in video generation [7]. Additionally, the OneTo3D method combines rapid implicit rendering with precise explicit control, allowing for dynamic and editable 3D representations [8].

Collectively, these principles and mechanisms emphasize the transformative role of diffusion models in video generation. The integration of probabilistic refinement, spatial-temporal correlations, and sophisticated attention mechanisms significantly enhances the performance and adaptability of diffusion models. This advancement not only improves the quality of video generation but also facilitates the development of more refined and user-friendly methodologies for creating videos, as evidenced by various innovative approaches that effectively manage video structure, content fidelity, and temporal consistency. [53, 54, 55, 4]

(a) Effect of Weight w on Image Quality[56]



(b) A diagram illustrating a generative model for image-to-image translation[57]



(c) Dynamic Dynamics: A Comprehensive Approach to Video Dynamics Analysis[23]

Figure 5: Examples of Principles and Mechanisms

As shown in Figure 5, The exploration of diffusion models and their underlying principles and mechanisms is vividly illustrated through a series of examples that highlight different facets of these models. The first example, "Effect of Weight w on Image Quality," demonstrates how varying a parameter, specifically the weight w, impacts the quality of images produced by a diffusion model. Displayed in a grid format, this example provides a comparative analysis of images of cars, with each row representing a different weight value ranging from $w = 0$ to $w = 4$. This visual representation aids in understanding how parameter adjustments can influence the output quality. The second example, "A diagram illustrating a generative model for image-to-image translation," delves into the mechanics of a machine learning model designed to transform input images into modified outputs. This diagram focuses on a specific task where an initial image, such as a photograph of a dog, is processed through an encoder and other components to produce a new image with subtle alterations. Finally, the example titled "Dynamic Dynamics: A Comprehensive Approach to Video Dynamics Analysis" offers a detailed overview of video dynamics, categorizing the analysis into inter-frame, inter-segment, and video-level dynamics. This comprehensive approach highlights the various factors and levels of analysis, such as instant and scene dynamics, that contribute to understanding the overall dynamics of video content. Together, these examples provide a multifaceted view of diffusion models, showcasing their versatility and the intricate mechanisms that drive their functionality. [**?**]meng2023distillation,mukherjee2024rissoleparameterefficientdiffusionmodels,liao2024evaluation)

5.2 Innovations in Diffusion Models

Recent advancements in diffusion models have introduced several innovative techniques that enhance their efficiency and broaden their applicability in video generation. The RISSOLE model exemplifies this progress by employing block-wise generation combined with retrieval-augmented guidance, ensuring coherence across generated blocks and optimizing the generative process [57]. This approach highlights the potential for maintaining high-quality outputs while managing computational resources effectively.

DEIS demonstrates significant innovations by achieving state-of-the-art sampling performance with limited function evaluations, addressing the efficiency bottlenecks commonly associated with diffusion models [58]. This advancement facilitates faster sampling processes, making diffusion models more practical for real-world applications.

Despite these innovations, diffusion models still face challenges related to computational efficiency during inference, often requiring numerous evaluation steps to generate samples [34]. Addressing these challenges is crucial for enhancing the applicability of diffusion models in various contexts.

InstructVideo introduces Segmental Video Reward (SegVR) and Temporally Attenuated Reward (TAR), which enhance fine-tuning efficiency and video quality, distinguishing it from traditional methods [36]. These techniques improve the alignment of generated videos with user preferences, ensuring higher quality outputs.

EfficientDM achieves quantization-aware training (QAT) level performance with post-training quantization (PTQ) level efficiency, utilizing a data-free approach to fine-tuning that significantly reduces computational demands [39]. This innovation is particularly beneficial in resource-constrained environments, where maintaining high-quality video generation is essential.

CONTEXTDIFF leverages cross-modal context to enhance the learning capacity of diffusion models, allowing for more nuanced and contextually relevant video generation [32]. This integration of multimodal data interactions underscores the versatility of diffusion models in handling complex generative tasks.

Additionally, the introduction of a benchmark for evaluating the effectiveness of state-of-the-art detectors in distinguishing synthetic images generated by diffusion models from real images, particularly in challenging conditions, provides valuable insights into the capabilities and limitations of current diffusion techniques [59].

The recent innovations in diffusion model techniques, such as structure and content-guided video synthesis, high-definition video generation, and advanced text-to-video performance enhancements, underscore the rapid evolution of this field. These advancements facilitate more efficient and flexible video generation solutions, allowing for improved control over content fidelity, temporal consistency, and artistic expression, ultimately leading to higher-quality outputs that meet diverse user needs in video editing and creation. [50, 21, 60, 61, 4]. The continuous exploration of novel architectures and methodologies is essential for overcoming the challenges of video generation and enhancing the capabilities of generative models.

5.3 Applications and Case Studies

Diffusion models have demonstrated significant versatility and efficiency in real-world video generation applications, effectively capturing complex spatial-temporal dynamics. The Latent Motion Diffusion Model (LaMD) exemplifies this by generating high-quality videos across diverse motion scenarios, achieving state-of-the-art performance in both image-to-video and text-image-to-video tasks [15]. This capability underscores the potential of diffusion models to integrate visual and textual data seamlessly, producing coherent video content.

In the realm of human motion generation, the MDM framework excels by balancing quality and resource efficiency, establishing new benchmarks across various motion generation tasks [62]. This achievement highlights the effectiveness of diffusion models in creating realistic and dynamic human movements, which are crucial for applications in animation and virtual reality.

The VIDM framework has been rigorously evaluated on datasets such as UCF-101, TaiChi-HD, Sky Time-lapse, and CLEVRER, utilizing Fréchet Video Distance (FVD) scores and visual quality assessments to benchmark against existing state-of-the-art methods [63]. These evaluations confirm the superior performance of diffusion models in generating visually appealing and temporally coherent video sequences.

Diffusion transformers have further enhanced these models' applications by effectively capturing spatial-temporal dependencies, thereby improving learning efficiency [64]. This advancement is particularly beneficial in scenarios requiring the synthesis of complex video content from minimal input data.

In practical terms, the shortcut MCMC sampling method has been evaluated for its performance in accelerating inference processes within diffusion models, demonstrating significant improvements over baseline methods across varying numbers of inference steps [65]. This innovation is crucial for real-time video generation applications where speed is of the essence.

Moreover, the DeepCache approach presents a novel method for accelerating diffusion models by leveraging the similarity in high-level features between denoising steps, thereby optimizing computational efficiency [66]. This technique is particularly advantageous in resource-constrained environments.

The evaluation of VIDIM on curated datasets like Davis-7 and UCF1017, which contain videos with large and ambiguous motion, further illustrates the capability of diffusion models in handling complex motion interpolation tasks [67]. This application is vital for enhancing video quality in post-production and editing processes.

Additionally, the Infusion method demonstrates significant improvements in video inpainting, particularly for dynamic textures, achieving state-of-the-art results with a lightweight model [1]. The CONTEXTDIFF model showcases state-of-the-art performance in text-to-image generation and text-to-video editing tasks, significantly enhancing the semantic alignment between text conditions and generated samples [32].

Experiments with PixelMan on benchmark datasets COCOEE and ReS reveal its effectiveness in consistent object editing, outperforming state-of-the-art methods [68]. Furthermore, the LaMoR dataset introduces a new benchmark for evaluating high-resolution frame interpolation methods, offering a valuable resource for assessing the capabilities of diffusion models in this domain [69].

Collectively, these case studies and applications demonstrate the transformative impact of diffusion models in real-world video generation, offering efficient and high-quality solutions across various domains. The continuous advancements in diffusion model techniques are significantly enhancing their versatility and effectiveness, leading to the development of more sophisticated and user-friendly methodologies for video generation. Recent innovations, such as structure and content-guided models, enable precise video editing based on descriptive prompts while maintaining visual consistency and temporal coherence. Additionally, methods like VideoElevator improve the quality of text-to-video generation by refining temporal motion and enhancing spatial details, thereby addressing previous limitations in frame quality and text alignment. As a result, these advancements are not only streamlining the video editing workflow but also expanding creative possibilities in various fields, including animation, advertising, and educational content creation. [60, 54, 50, 4]

6 Video Synthesis

Video synthesis involves diverse methodologies aimed at generating coherent and realistic video content. This section explores foundational techniques for constructing video sequences essential for producing high-quality outputs that align with specific narratives and contexts. The subsequent subsection highlights innovative approaches that enhance video sequence construction, emphasizing the synergy between technology and creativity in this evolving field.

6.1 Techniques for Video Sequence Construction

Constructing video sequences involves synthesizing coherent content from existing data, employing innovative methodologies to enhance realism and contextual relevance. Techniques such as Control-A-Video utilize control maps, like edge and depth maps, to guide video generation, thus improving the precision and quality of synthesized sequences [70]. This incorporation of structural information is vital for achieving detailed and controlled outputs.

Collaborative Video Diffusion (CVD) demonstrates the scalability and flexibility of modern techniques by enabling multiple video generations from a single trained model [71]. Similarly, StyleGAN-V generates videos as continuous-time signals, incorporating motion information through positional embeddings to ensure temporal coherence and smooth transitions [49].

Datasets like MIRAData, which include structured captions and diverse content such as 3D engine-rendered scenes and human motion, enhance the descriptive depth and contextual accuracy of

synthesized videos [72]. The integration of rich metadata supports nuanced video synthesis aligned with intended narratives.

Emerging standards, highlighted in the MPEG survey, provide compact descriptors for visual search and analysis, ensuring interoperability and consistency across platforms [14]. These standards are crucial for maintaining video sequence quality and coherence.

Innovative methods like ActAnywhere dynamically integrate subjects with synthesized backgrounds based on input segmentation and condition frames, allowing for interactive and contextually relevant content [73]. MoCoGAN combines fixed content vectors with dynamically generated motion vectors to construct video frames, effectively merging static and dynamic elements in synthesis [74].

The ShareGPT4Video dataset offers high-quality video-caption pairs, serving as a valuable resource for developing techniques that create video sequences with rich world knowledge and detailed temporal descriptions [24]. This dataset supports generating visually coherent and contextually enriched videos.

In virtual try-ons, methods like ViViD synthesize videos where the target person wears a specified garment while preserving the original video's context, showcasing the adaptability of video synthesis techniques in fashion and retail [37]. Additionally, generating novel views through sampling distributions consistent with input images and camera parameters exemplifies the ability to create new perspectives while maintaining original content integrity [29].

These techniques illustrate diverse methodologies in video sequence construction, each contributing to advancing realistic and contextually aligned video synthesis. Progress is driven by sophisticated control mechanisms, extensive datasets, and cutting-edge modeling techniques. Recent innovations, including video-to-video synthesis frameworks and compositional text-to-video generation models, expand the horizons of video content creation, facilitating the synthesis of intricate scenes and the manipulation of multiple objects over time. This evolution enhances video content creators' capabilities across fields like computer vision, robotics, and entertainment [22, 6, 26, 25, 4].

6.2 Challenges in Realistic and Coherent Video Synthesis

Generating realistic and coherent video content presents significant challenges, primarily due to the complexity of capturing precise temporal continuity and motion dynamics. High-resolution video synthesis often requires extensive fine-tuning, leading to issues such as pattern repetition and desaturation, degrading visual quality [75]. Maintaining quality and stability in long video sequences remains challenging, as current methods frequently struggle to preserve visual fidelity over extended durations [49].

In scenarios requiring intricate spatial conditioning, such as laparoscopic video generation, achieving precision in tool movement and spatial alignment is particularly challenging [76]. The reliance on high-quality training data further limits the applicability of synthesis methods, as insufficient data can result in inconsistencies and visual artifacts.

Despite advancements in camera motion control, achieving high image quality and rich content remains problematic, especially when models are trained on smaller datasets, leading to deformations and inconsistencies [77]. Techniques like MCVD often struggle with generating longer sequences, resulting in blurry or inconsistent outputs, particularly in unconditional generation settings [78].

Ethical concerns regarding the misuse of generated content pose additional challenges, necessitating careful consideration in developing and deploying video synthesis technologies [79]. Achieving controllability and consistency across frames in high-resolution videos remains formidable, as current methods often fail to maintain visual quality throughout the video [80].

The regeneration quality of video content is frequently constrained by the base model, which may introduce artifacts that compromise overall coherence [19]. Furthermore, maintaining object consistency during significant movements is limited by the underlying model's capabilities, affecting the quality of generated videos [81].

Finally, the Diffusion4D framework emphasizes the need for higher resolution and longer temporal sequences to enhance the fidelity of generated 4D content, highlighting ongoing challenges in achieving realistic and coherent video synthesis [17]. These multifaceted challenges necessitate continued research and innovation to advance the field of video synthesis.

6.3 Innovations in Video Synthesis

Recent innovations in video synthesis have introduced advancements enhancing the realism and coherence of generated content. Notably, integrating continuous motion codes and a holistic discriminator in StyleGAN-V aggregates temporal information, improving temporal coherence and quality [49].

There is an increasing focus on optimizing video synthesis efficiency. Research into memory-efficient optimization methods and exploring efficient diffusion models for video and 3D generation tasks represent promising directions aimed at reducing computational demands while maintaining high-quality outputs [39].

Advancements in video synthesis techniques, evidenced by innovations like search-based generation pipelines and compositional frameworks such as VideoTetris, underscore ongoing efforts to enhance realism and computational efficiency. These approaches improve the accuracy of motion dynamics and object interactions in text-to-video synthesis while implementing efficient algorithms that reduce the need for extensive data training, making the generation process more accessible and effective [25, 22, 23]. Continuous exploration of novel methodologies and optimization strategies is crucial for advancing the field and enabling sophisticated video content creation.

7 Generative Models

7.1 Learning Patterns from Large Datasets

Generative models have revolutionized video generation by effectively learning intricate patterns from extensive datasets. Utilizing frameworks like diffusion and autoregressive models, they capture complex temporal and spatial dynamics, enabling the creation of realistic and coherent video content across diverse scenarios [15]. The Latent Video Diffusion Model (LVDM) exemplifies this by optimizing computational efficiency in a low-dimensional latent space while maintaining high fidelity [2].

Large datasets empower these models to discern detailed patterns, including motion dynamics and scene transitions, crucial for high-quality video sequences. This capability addresses challenges such as motion blur and artifacts, enhancing temporal coherence and spatial consistency [7]. The adaptability of generative models to new patterns ensures their ongoing relevance in video generation, catering to the rising demand for realistic and immersive content in applications like entertainment and virtual reality [49].

Models like VideoGPT and diffusion architectures demonstrate the ability to learn complex patterns from large datasets, improving the quality of synthesized content by producing high-fidelity outputs aligned with textual prompts. This enhances generation efficiency and advances temporal consistency and dynamic representation [23, 82, 22, 51, 4]. The evolution of these models, leveraging extensive data resources, promises further enhancements in video generation technologies, paving the way for sophisticated and accessible content creation.

7.2 Model Complexity vs. Generation Efficiency

Balancing model complexity and generation efficiency is crucial in video generation, influencing the feasibility and scalability of generative models. While complex models yield high-quality video content, they often require substantial computational resources, limiting their use in resource-constrained environments. The UVCG model exemplifies a balanced approach, leveraging temporal consistency to map continuous inputs to misaligned outputs, enhancing efficiency without compromising quality [83].

Advanced architectures with attention mechanisms and hierarchical structures excel in capturing intricate dynamics. However, recent advancements in text-to-video (T2V) models, such as Make-A-Video and search-based pipelines, while improving video quality and realism, increase computational demands, challenging real-time generation [23, 13, 6, 25, 84]. Optimizing model complexity is essential for efficient deployment across applications.

Efficient generative models minimize parameters and computational steps through techniques like model pruning, quantization, and low-dimensional latent spaces. For instance, the Attention-driven

Training-free Efficient Diffusion Model (AT-EDM) reduces computational load by 38.8% without retraining, while the EfficientDM framework combines low-rank adapters with model weights for efficient quantization [57, 85, 27, 45, 39]. These strategies are vital for deploying complex models in resource-limited scenarios.

Balancing complexity and efficiency is key in designing generative models for video generation. By optimizing dynamics and content fidelity, researchers can develop advanced T2V models that deliver high-quality video content with reduced computational demands, broadening the applicability of T2V technologies across platforms [60, 23, 6, 84].

7.3 Innovations in Generative Video Models

Recent advancements in generative video models have introduced innovative techniques that significantly enhance video content creation quality and efficiency. Diffusion models have proven exceptional in producing high-fidelity and temporally consistent video sequences. Innovations like structure and content-guided video diffusion models allow user modifications based on textual descriptions while maintaining structural integrity. Models such as VideoTetris and CogVideoX address complexities in generating videos with multiple objects and dynamic scenes, ensuring coherent narratives and improved alignment between text prompts and video content [22, 53, 21, 86, 4].

The Latent Motion Diffusion Model (LaMD) enhances generative video models by capturing dynamic interactions and complex motion patterns within sequences, optimizing computational efficiency while preserving video fidelity [15]. Hierarchical modeling and noise conditioning techniques further bolster model robustness and adaptability [2].

Attention mechanisms have been pivotal in advancing generative video models. Techniques like MotionCom leverage learned motion dynamics to integrate objects into video sequences seamlessly, ensuring realism and contextual relevance [3]. Attention-based models improve video quality by managing spatial-temporal correlations effectively.

Efficient quantization strategies, exemplified by the EfficientDM framework, address resource constraints in generative video modeling. This framework achieves quantization-aware training (QAT) level performance with post-training quantization (PTQ) level efficiency, facilitating high-quality video generation in resource-limited environments [39].

The ongoing evolution of generative video models, marked by innovative frameworks and evaluation protocols, emphasizes video dynamics essential for visual authenticity and coherence with text prompts. The DEVIL evaluation protocol highlights the importance of dynamics in assessing T2V generation models, while frameworks like VideoTetris and DreamVideo tackle compositional generation challenges and customized video creation [23, 22, 53, 51, 4]. Exploring novel architectures and methodologies promises to elevate generative models' capabilities, paving the way for sophisticated and accessible video content creation.

8 Conclusion

8.1 Future Directions

Future research in efficient video generation is poised to focus on optimizing architectural designs and training methodologies to enhance performance while minimizing computational demands. Emphasis will be placed on creating models capable of generating extended videos with multiple scenes and events, while addressing social biases in training data to ensure equitable outcomes. Enhancing synthesis speed and audio-lip synchronization across various audio inputs remains crucial, alongside addressing potential misuse of video generation technologies.

Incorporating additional conditioning data, such as facial landmarks and pose estimates, is crucial for improving the fidelity of generated results and overcoming current limitations in video synthesis. Moreover, refining bounding box extraction methods and model architectures is essential for bolstering efficiency and fidelity in object-centric video generation.

Within audio generation, the development of lighter and faster diffusion models is imperative to improve efficiency and reduce computational overhead, thereby facilitating more accessible audio-visual content creation. Advancements in realistic channel models and diffusion model architectures

are vital for enhancing video generation capabilities, alongside comprehensive analyses of sample complexity and training duration.

Exploring multi-condition approaches to augment video generation quality and the capacity to produce extended videos with richer semantic content are pivotal areas for future exploration. Progress in models that enhance efficiency without compromising quality, through innovative architectures and computational techniques, is essential for the field's advancement.

In dynamic scene generation, future research should prioritize improving accuracy in camera pose and object motion, while investigating advanced techniques for 3D reconstruction. Enhancing the capabilities of models like OneTo3D for more complex dynamic scenes and optimizing computational efficiency are critical areas for development.

Finally, expanding the scope of dynamics grades and evaluating a wider array of text-to-video models will be necessary to validate the effectiveness of current benchmarks, ensuring comprehensive assessments of video generation models. These future research directions collectively underscore ongoing efforts to advance efficient video generation, with a focus on adaptability, control, and computational efficiency to meet the evolving demands of various applications.

References

[1] Nicolas Cherel, Andrés Almansa, Yann Gousseau, and Alasdair Newson. Infusion: internal diffusion for inpainting of dynamic textures and complex motion, 2024.

[2] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

[3] Weijing Tao, Xiaofeng Yang, Miaomiao Cui, and Guosheng Lin. Motioncom: Automatic and motion-aware image composition with llm and video diffusion prior, 2024.

[4] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7346–7356, 2023.

[5] Wilson Yan, Ryo Okumura, Stephen James, and Pieter Abbeel. Patch-based object-centric transformers for efficient video generation, 2022.

[6] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[7] Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, László Jeni, Sergey Tulyakov, and Hsin-Ying Lee. 4real: Towards photorealistic 4d scene generation via video diffusion models. *Advances in Neural Information Processing Systems*, 37:45256–45280, 2024.

[8] Jinwei Lin. Oneto3d: One image to re-editable dynamic 3d model and video generation, 2024.

[9] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models, 2023.

[10] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation, 2023.

[11] Anwaar Ulhaq and Naveed Akhtar. Efficient diffusion models for vision: A survey, 2024.

[12] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[13] Ashkan Taghipour, Morteza Ghahremani, Mohammed Bennamoun, Aref Miri Rekavandi, Zinuo Li, Hamid Laga, and Farid Boussaid. Faster image2video generation: A closer look at clip image embedding's impact on spatio-temporal cross-attentions, 2024.

[14] Lingyu Duan, Jiaying Liu, Wenhan Yang, Tiejun Huang, and Wen Gao. Video coding for machines: A paradigm of collaborative compression and intelligent analytics. *IEEE Transactions on Image Processing*, 29:8680–8695, 2020.

[15] Yaosi Hu, Zhenzhong Chen, and Chong Luo. Lamd: Latent motion diffusion for video generation, 2023.

[16] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[17] Hanwen Liang, Yuyang Yin, Dejia Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N. Plataniotis, Yao Zhao, and Yunchao Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models, 2024.

[18] Haoran Lang, Yuxuan Ge, and Zheng Tian. S2dm: Sector-shaped diffusion models for video generation, 2024.

[19] Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control, 2024.

[20] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling, 2024.

[21] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022.

[22] Ye Tian, Ling Yang, Haotian Yang, Yuan Gao, Yufan Deng, Jingmin Chen, Xintao Wang, Zhaochen Yu, Xin Tao, Pengfei Wan, et al. Videotetris: Towards compositional text-to-video generation. *arXiv preprint arXiv:2406.04277*, 2024.

[23] Mingxiang Liao, Qixiang Ye, Wangmeng Zuo, Fang Wan, Tianyu Wang, Yuzhong Zhao, Jingdong Wang, Xinyu Zhang, et al. Evaluation of text-to-video generation models: A dynamics perspective. *Advances in Neural Information Processing Systems*, 37:109790–109816, 2024.

[24] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024.

[25] Haoran Cheng, Liang Peng, Linxuan Xia, Yuepeng Hu, Hengjia Li, Qinglin Lu, Xiaofei He, and Boxi Wu. Searching priors makes text-to-video synthesis better, 2024.

[26] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

[27] Luís Arandas, Mick Grierson, and Miguel Carvalhais. Antagonising explanation and revealing bias directly through sequencing and multimodal inference, 2023.

[28] Chen Rao, Guangyuan Li, Zehua Lan, Jiakai Sun, Junsheng Luan, Wei Xing, Lei Zhao, Huaizhong Lin, Jianfeng Dong, and Dalong Zhang. Rethinking video deblurring with wavelet-aware dynamic transformer and diffusion model, 2024.

[29] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models, 2023.

[30] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[31] Shiyin Dong, Mingrui Zhu, Kun Cheng, Nannan Wang, and Xinbo Gao. Bridging generative and discriminative models for unified visual perception with diffusion priors, 2024.

[32] Ling Yang, Zhilong Zhang, Zhaochen Yu, Jingwei Liu, Minkai Xu, Stefano Ermon, and Bin Cui. Contextualized diffusion models for text-guided image and video generation, 2024.

[33] Zichen Miao, Zhengyuan Yang, Kevin Lin, Ze Wang, Zicheng Liu, Lijuan Wang, and Qiang Qiu. Tuning timestep-distilled diffusion model using pairwise sample optimization, 2025.

[34] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

[35] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[36] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback, 2023.

[37] Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. Vivid: Video virtual try-on using diffusion models, 2024.

[38] Yiheng Huang, Hui Yang, Chuanchen Luo, Yuxi Wang, Shibiao Xu, Zhaoxiang Zhang, Man Zhang, and Junran Peng. Stablemofusion: Towards robust and efficient diffusion-based motion generation framework, 2024.

[39] Yefei He, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Efficientdm: Efficient quantization-aware fine-tuning of low-bit diffusion models, 2024.

[40] Muah Kim, Rick Fritschek, and Rafael F. Schaefer. Learning end-to-end channel coding with diffusion models, 2023.

[41] Wenhao Li, Yichao Cao, Xiu Su, Xi Lin, Shan You, Mingkai Zheng, Yi Chen, and Chang Xu. Training-free long video generation with chain of diffusion model experts, 2024.

[42] Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller, 2024.

[43] Zhixing Zhang, Yanyu Li, Yushu Wu, Yanwu Xu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Dimitris Metaxas, Sergey Tulyakov, and Jian Ren. Sf-v: Single forward video generation model, 2024.

[44] Tim Kaiser, Nikolas Adaloglou, and Markus Kollmann. The unreasonable effectiveness of guidance for diffusion models, 2024.

[45] Huijie Zhang, Yifu Lu, Ismail Alkhouri, Saiprasad Ravishankar, Dogyoon Song, and Qing Qu. Improving efficiency of diffusion models via multi-stage framework and tailored multi-decoder architectures, 2024.

[46] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022.

[47] Matthew Tivnan, Jacopo Teneggi, Tzu-Cheng Lee, Ruoqiao Zhang, Kirsten Boedeker, Liang Cai, Grace J. Gang, Jeremias Sulam, and J. Webster Stayman. Fourier diffusion models: A method to control mtf and nps in score-based stochastic image generation, 2023.

[48] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities, 2024.

[49] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3626–3636, 2022.

[50] Elijah Miller, Thomas Dupont, and Mingming Wang. Enhanced creativity and ideation through stable video synthesis, 2024.

[51] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion, 2023.

[52] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation, 2023.

[53] Hyelin Nam, Jaemin Kim, Dohun Lee, and Jong Chul Ye. Optical-flow guided prompt optimization for coherent video generation, 2024.

[54] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023.

[55] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18456–18466, 2023.

[56] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023.

[57] Avideep Mukherjee, Soumya Banerjee, Piyush Rai, and Vinay P. Namboodiri. Rissole: Parameter-efficient diffusion models via block-wise generation and retrieval-guidance, 2024.

[58] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

[59] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models, 2022.

[60] Yabo Zhang, Yuxiang Wei, Xianhui Lin, Zheng Hui, Peiran Ren, Xuansong Xie, Xiangyang Ji, and Wangmeng Zuo. Videoelevator: Elevating video generation quality with versatile text-to-image diffusion models, 2024.

[61] Nisha Huang, Yuxin Zhang, and Weiming Dong. Style-a-video: Agile diffusion for arbitrary text-based video style transfer, 2023.

[62] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[63] Vidm: Video implicit diffusion models.

[64] Hengyu Fu, Zehao Dou, Jiawei Guo, Mengdi Wang, and Minshuo Chen. Diffusion transformer captures spatial-temporal dependencies: A theory for gaussian process data, 2025.

[65] Gang Chen. Speed up the inference of diffusion models via shortcut mcmc sampling, 2022.

[66] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free, 2023.

[67] Siddhant Jain, Daniel Watson, Eric Tabellion, Aleksander Hołyński, Ben Poole, and Janne Kontkanen. Video interpolation with diffusion models, 2024.

[68] Liyao Jiang, Negar Hassanpour, Mohammad Salameh, Mohammadreza Samadi, Jiao He, Fengyu Sun, and Di Niu. Pixelman: Consistent object editing with diffusion models via pixel manipulation and generation, 2025.

[69] Junhwa Hur, Charles Herrmann, Saurabh Saxena, Janne Kontkanen, Wei-Sheng Lai, Yichang Shih, Michael Rubinstein, David J. Fleet, and Deqing Sun. High-resolution frame interpolation with patch-based cascaded diffusion, 2024.

[70] Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning, 2024.

[71] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *Advances in Neural Information Processing Systems*, 37:16240–16271, 2024.

[72] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.

[73] Boxiao Pan, Zhan Xu, Chun-Hao Huang, Krishna Kumar Singh, Yang Zhou, Leonidas J Guibas, and Jimei Yang. Actanywhere: Subject-aware video background generation. *Advances in Neural Information Processing Systems*, 37:29754–29776, 2024.

[74] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.

[75] Lanqing Guo, Yingqing He, Haoxin Chen, Menghan Xia, Xiaodong Cun, Yufei Wang, Siyu Huang, Yong Zhang, Xintao Wang, Qifeng Chen, Ying Shan, and Bihan Wen. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation, 2024.

[76] Ivan Iliash, Simeon Allmendinger, Felix Meissen, Niklas Kühl, and Daniel Rückert. Interactive generation of laparoscopic videos with diffusion models, 2024.

[77] Soon Yau Cheong, Duygu Ceylan, Armin Mustafa, Andrew Gilbert, and Chun-Hao Paul Huang. Boosting camera motion control for video diffusion transformers, 2024.

[78] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022.

[79] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models, 2023.

[80] Zhongjie Duan, Chengyu Wang, Cen Chen, Weining Qian, and Jun Huang. Diffutoon: High-resolution editable toon shading via diffusion models, 2024.

[81] Haonan Qiu, Zhaoxi Chen, Zhouxia Wang, Yingqing He, Menghan Xia, and Ziwei Liu. Freetraj: Tuning-free trajectory control in video diffusion models, 2024.

[82] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[83] KaiZhou Li, Jindong Gu, Xinchun Yu, Junjie Cao, Yansong Tang, and Xiao-Ping Zhang. Uvcg: Leveraging temporal consistency for universal video protection, 2024.

[84] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

[85] Hongjie Wang, Difan Liu, Yan Kang, Yijun Li, Zhe Lin, Niraj K. Jha, and Yuchen Liu. Attention-driven training-free efficiency enhancement of diffusion models, 2024.

[86] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.