Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 10, No. 02, June 2024.

TEXT TO VIDEO USING GANS AND DIFFUSION MODELS

Nikita Singhal, Praval Pratap Singh, Nikhil Singh, Mahipal Singh and Harsimrat Singh

(Received: 21-Feb.-2024, Revised: 5-Apr.-2024, Accepted: 24-Apr.-2024)

ABSTRACT

The challenging endeavour of text-to-video creation requires transforming text descriptions into realistic and cohesive videos. This field of study has made substantial progress in recent years, with the development of diffusion models and generative adversarial networks (GANs). This study examines the most modern text-to-video generation models, as well as the various steps involved in text-to-video generation, including temporal coherence, video generation and text encoding. We additionally emphasise the challenges involved with text-to-video generation, as well as recent advances to overcome these issues. The most frequently used datasets and metrics in this field are also analyzed and reviewed.

KEYWORDS

Text-to-video, Coherence, GAN, Diffusion.

1. INTRODUCTION

Video generation has grown dramatically in recent years, gaining popularity due to its various advantages and applications in a variety of sectors, including marketing, branding, content development, artificial video-dataset generation and so on. The objective of this article is to review and compare various text-to-video (T2V) generation approaches. Our goal is to investigate various models across various stages. Very few articles investigated video generation in depth. Furthermore, as new approaches are discovered, it becomes necessary to compare them in order to identify the limitations and constraints of existing techniques, which may then be used by other researchers for future study and enhancements. Table 1 compares the proposed survey study to existing T2V survey studies, including their features and limitations.

T2V is the next step after text-to-image (T2I). Like T2I, T2V began with the use of GAN[3] models and proceeded to the use of different techniques, the most common of which is diffusion, which allows us to use previously existing text to image models. A lot of study has been done in the text-to-image field, since it is used in T2V in diffusion video models.

In the beginning, GANs, which were excellent at generating images at the time, were used to generate images from text. However, stable diffusion has grabbed the lead in producing images of excellent quality in recent years. Different methodologies and tactics for addressing additional concerns, such as temporal and spatial consistency, were considered. Furthermore, the metrics used to evaluate text-to-video models have changed and novel metrics, such as FVD [4], have been established to provide further insight into text-to-video models. Some well-known models, such as Text2VideoZero [5] and Hotshot-XL [6], are also evaluated in terms of how well they perform using an FVD matrix.

The rest of the paper is organized as follows. Section 2 summarizes the various stages and approaches employed in T2V. In Section 3, we discussed the most often used datasets in T2V. In Section 4, we reviewed numerous metrics for evaluating T2V performance. In Section 5, we discussed open challenges and directions for future research and in Section 6, we concluded the work.

2. LITERATURE SURVEY

2.1 Video Generation

Video generation, a dynamic field at the intersection of artificial intelligence and multimedia, encompasses a spectrum of techniques dedicated to converting conditional and unconditional information into captivating visual content. The process involves a thoughtful blend of natural language

N. Singhal, P. P. Singh, N. Singh, M. Singh and H. Singh are with Department of Computer Engineering, Army Institute of Technology, Pune, India. Emails: [ngupta,pravalsingh_20250, nikhilsingh_20229,mahipalsingh_20134,harsimatsingh_20060]@aitpune.edu.in

processing (NLP), machine learning and creative design principles.

Table 1. Comparison of existing studies with proposed studies.

Study & Year	**Advantages**	**Limitations**
[1], 2023	• Provides an overview of existing literature of T2I and T2V AI generation • Theoretical comparison of different T2I and T2V models	• Performance evaluation is not conducted • Does not describe the processes involved in T2I or T2V generation
[2], 2023	• Comprehensive coverage-covering domains : video generation, editing and understanding • In-depth examination of the diffusion-model applications in the context of video • Conducted Performance evaluation	• Does not explain internal processes involved in T2V generation
Proposed Review	• Comprehensive coverage of video generation using textual input • Discussed the internal processes involved in T2V generation (including T2I, cross frame attention, motion dynamics and frame interpolation) • Conducted performance evaluation of various models using FVD score • Evaluated FVD score for models that were not included in [2]	

As video generation continues to evolve, researchers explore novel ways to dynamically generate scenes, integrate user feedback and enhance content creation through adaptive systems. This fusion of technology and creativity not only automates the process of video production, but also opens new frontiers for personalized and engaging multimedia experiences. Whether used in education, entertainment or communication, video generation stands as a testament to the ever-expanding capabilities of AI in transforming textual narratives into visually compelling stories.

Zhen Xing et al. [2] categorized video generation into four categories: Text-to-video generation, video-generation using different conditions, unconditional video generation and video-completion. In the proposed survey, we will delve into a comprehensive exploration of text-to-video generation thoroughly examining the various steps involved in the text-to-video generation process.

2.2 Text-to-video Generation

Video generation using GANs and diffusion models represents a cutting-edge approach in the realm of artificial intelligence and computer vision. GANs, pioneered by Ian Goodfellow et al. [3], consist of two neural networks, the generator and the discriminator, engaged in a game-like scenario, where the generator strives to create realistic data (in this case, video frames), while the discriminator aims to differentiate between real and generated data. This adversarial training process leads to the generator producing increasingly realistic video sequences over time. The entire system optimizes a function denoted as in Equation (1):

$$V(D,G) = E_{X\sim p_{\text{data}}(x)}[\log D(x)] + E_{Z\sim pz\,(z)}[\log(1-D(G(z)))] \quad (1)$$

where D(x) denotes the output of the discriminator for real data (x), G(z)) denotes the output value of the generator for latent vector (z), Ex∼pdata(x)[log D(x)] is the desired output value of the discriminator for actual data (x), Ez∼pz (z)[log(1 − D(G(z)))] denotes the desired output value of the discriminator for generated data (G(z)), x is derived using data distribution $p_{data}(x)$ and z is derived using the latent space distribution pz (z). Video GAN generators employ four primary strategies to effectively generate realistic video sequences. Firstly, they often utilize a hybrid approach that combines Recurrent Neural Network (RNN) architectures using 2D CNN to handle both temporal and spatial information. Secondly, some models opt for 3D convolutional networks instead of 2D ones to directly capture spatiotemporal

features from video data. Additionally, inspired by progressive growing GAN architecture, video GANs implement a coarse-to-fine strategy, refining generated data progressively for enhanced output. Lastly, they may adopt a two-stream architecture, with parallel streams specialized in processing different aspects of video data, aiding in capturing spatial and temporal features effectively. For the discriminator, strategies such as using a two-stream architecture or a 3D convolutional network are employed to distinguish between real and generated video data based on their effectiveness. These strategies collectively contribute to the advancement of video generation techniques. In recent years, researchers have integrated diffusion models into the video generation process to optimize the quality and realism of generated content. Diffusion models, inspired by the concept of Brownian motion, simulate the gradual spreading or diffusion of information or features across a data space. Diffusion models can capture long-range dependencies and temporal coherence, allowing for the creation of smoother and more natural-looking video sequences.

Text-to-video generation using diffusion models involves a series of interconnected steps orchestrated to transform textual prompts into coherent video sequences, as depicted in Figure 1. Initially, a text prompt is provided as input, which undergoes encoding by a text encoder to capture its semantic meaning, resulting in a fixed-length text embedding. Concurrently, a scheduler controls the diffusion process, modulating noise application to a latent image over successive time steps. This latent image represents the evolving state of video generation and serves as a canvas for subsequent transformations. Incorporating temporal information, timestamp embedding encodes frame sequences, facilitating coherent motion dynamics throughout the video-generation process. Alongside, text embedding, derived from the text prompt, as well as timestamp embedding, are concatenated and utilized by the decoder to synthesize each frame. The motion dynamics within the latent code are shaped by these embeddings, aligning the generated video with the provided text and ensuring smooth temporal progression. Integral to the process is the diffusion model or noise predictor, like U-NET, which models the conditional distribution of subsequent frames based on the current frame and noise level. Cross-frame attention mechanisms capture dependencies between frames, enabling the model to maintain coherence and consistency across the video sequence. Finally, frame-interpolation techniques may be employed to generate intermediate frames for smooth transitions, while background smoothing enhances visual quality and reduces artifacts, ensuring the fidelity of the generated video. Through this orchestrated flow, text-to-video generation using diffusion models seamlessly translates textual descriptions into visually compelling video content.

Diffusion models [7]-[8] acquire the ability to create data by progressively refining samples taken from a noise distribution. Gaussian diffusion models operate under the assumption of a forward noising process, where noise (ϵ) is gradually added to genuine data ($x0 \sim p_{data}$). The mathematical definition denoting the forward noising process is represented in Equation (2):

$$x_t = \sqrt{\gamma(t)}x_0 + \sqrt{1-\gamma(t)}\epsilon, \epsilon \sim N(0,I), t \in [0,1] \quad (2)$$

where $\gamma(t)$ represents a function that steadily decreases from 1 to 0 (referred to as the "noise schedule"). Diffusion models are trained for converse procedure, which counteracts the initial corruptions introduced during the forward process. The mathematical definition denoting the converse noising procedure is represented in Equation (3):

$$E_{x \sim p_{data}, t \sim U(0,1),\ \epsilon \sim N(0,I)}[\|y - f_\theta(x_t; c, t\|^2] \quad (3)$$

where $f\theta$ represents the denoiser model, which is defined by a neural network's parameters, conditioning information is donoted by c, such as textual prompts or class labels, while the target y can be arbitrary noise ϵ,and the denoised input x_0 or $v = \sqrt{1-\gamma(t)}\epsilon - \sqrt{\gamma(t)}x_0$. Combining GANs and diffusion models for video generation involves leveraging the strengths of both approaches. GANs excel at capturing high-frequency details and local structure in video frames, while diffusion models are effective at modeling long-term dependencies and global temporal coherence. By integrating these techniques, researchers have achieved significant advancements in generating high-resolution, photorealistic videos with coherent motion and semantic consistency. The synergy between GANs and diffusion models opens up new possibilities for applications, such as video synthesis, content creation and video editing. Furthermore, ongoing research in this field continues to push the boundaries of what is achievable, paving the way for more sophisticated and life-like video-generation systems in the future.

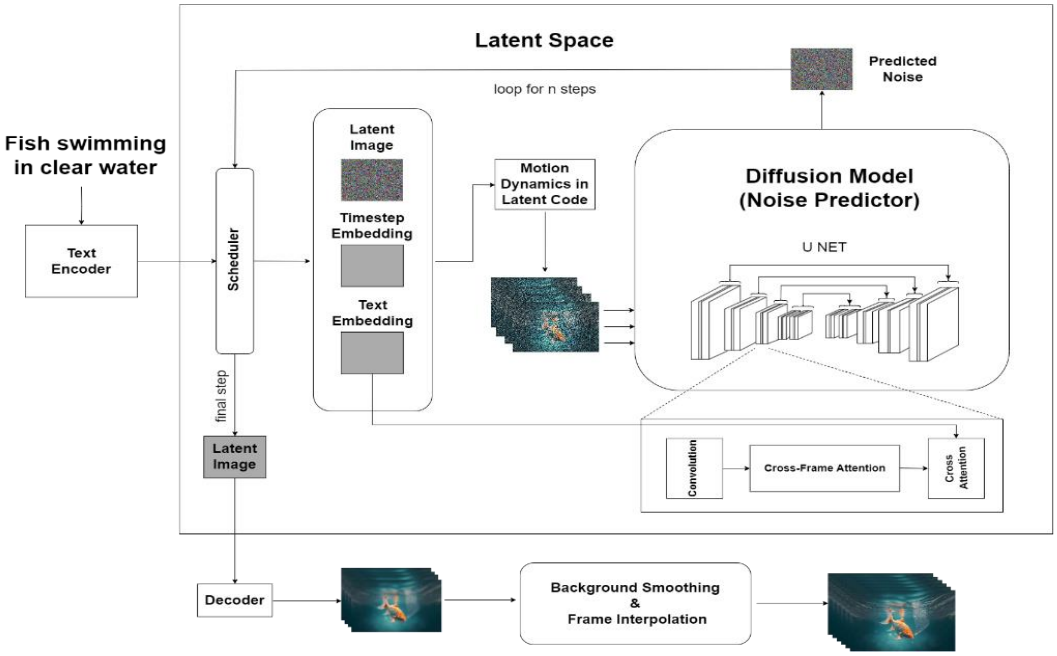

Figure 1. Architecture of a general text-to-video model using stable diffusion.

Video generation with text condition is further divided into two distinct categories: training-based and training-free T2V diffusion methods.

1. Training-based T2V Diffusion Methods: Diverse approaches are used for video synthesis in training- based text-to-video diffusion models, which improve quality by introducing novel training tactics and techniques. The most widely used training-based text-to-video diffusion models are listed here, along with some of their key features.

    - Make-A-Video, as described by Singer et al. [9], revolutionizes the process of learning visual- textual associations by leveraging paired image-text data and extracting video dynamics from unsupervised video datasets. This approach minimizes the need for extensive data-collection efforts, facilitating the generation of diverse and life-like videos through the integration of multiple super-resolution models and interpolation networks.
    - Imagen Video [10], an extension of the established T2I model Imagen [11], introduces a cascaded video-diffusion model consisting of seven interconnected sub-models. The effectiveness of training methodologies, such as classifier-free guidance, conditioning augmentation and v-parameterization, is validated, with additional benefits achieved through progressive distillation techniques aimed at enhancing sampling efficiency.
    - Show-1 [12], introduced by Zhang et al. (2023), innovates by combining pixel-based and latent- based diffusion models for T2V generation. This model operates across four distinct stages, each focusing on different aspects, such as key-frame generation, frame interpolation, super-resolution and latent super-resolution modules, thereby enhancing the overall video-synthesis process.
    - MagicVideo [13], developed by Zhou et al. (2022), employs the Latent Diffusion Model (LDM) to generate videos in latent space, effectively reducing computational overhead and accelerating processing speed. A frame-wise lightweight adaptor is introduced to align distributions, thereby improving temporal relationship modeling and the overall video quality.
    - Latent-Shift [14], as presented by An et al. (2023), prioritizes lightweight temporal modeling inspired by Temporal Shift Module (TSM). This approach involves channel shifting between adjacent frames within convolutional blocks, ensuring the retention of T2I capabilities while generating videos.
    - ModelScope [15], as described by Wang et al. (2023), integrates spatial-temporal convolution and attention mechanisms into the Latent Diffusion Model (LDM)

framework for T2V tasks. Leveraging a mixed training approach utilizing LAION and WebVid datasets, it serves as an open-source benchmark for T2V-synthesis methods.

- VideoFusion [16], proposed by Luo et al. (2023), addresses content redundancy and temporal correlations by decomposing the diffusion process using shared base noise and residual noise along the temporal axis for each frame. Two co-training networks are employed for noise decomposition, ensuring coherence in frame motion and improving the overall video-synthesis quality.

2. Training-free T2V diffusion methods: Training-free text-to-video (T2V) diffusion methods involve direct synthesis or utilize pre-existing models without dedicated training, bypassing explicit training processes.

    - Text2Video-Zero [5] uses a pre-trained text-to-image model for the purpose of video generation, incorporating a cross-attention mechanism and modifying latent code sampling to enhance motion dynamics.
    - DirecT2V [17] and Free-Bloom [18] employ large language models (LLMs) for frame-to-frame descriptions based on user prompts. DirecT2V uses dual-softmax filtering and value mapping for continuity between frames, while Free-Bloom introduces enhancements, like joint noise sampling and step-aware attention shifting.

2.3 Processes Involved in Text to Video Generation

In the intricate process of T2V generation, several key stages unfold. Initially, the system undertakes T2I generation, crafting a single frame that encapsulates the visual representation of the provided textual input. Subsequently, the model engages in cross-frame attention and motion dynamics, employing attention mechanisms to intricately link frames and model the dynamic motion inherent in the video sequence. This step ensures a coherent and realistic flow between frames. Finally, frame Interpolation comes into play, facilitating the creation of intermediate frames. This interpolation process enhances temporal continuity, contributing to the seamless generation of a cohesive and visually compelling video sequence. Together, these stages form a comprehensive pipeline for the transformation of text descriptions into dynamic and visually engaging video content.

2.3.1 Text-to-Image (Generation of a Single Frame)

T2I models serve as a foundational stage and point of entry for T2V models. Currently, there are various cutting-edge models based on stable diffusion and GANs. Table 2 provides a summary of popular studies in this field of study.

Generative Adversial Networks (GANs) [3] are unsupervised machine learning methods that function like supervised ones. Discriminators and generators are the two components of a GAN. While discriminators attempt to determine whether an image is real or not, generators attempt to create new images from the original dataset. In a zero-sum game, both players participate and the game ends when generators trick the discriminator more often than not. The two main advantages of GANs are their speed of inference and their ability to manipulate latent spaces to influence the synthesized outcome. A well-researched latent space in StyleGAN enables principled control over generated images. Diffusion models have made significant strides toward speeding up, but they are still far behind GANs, which only need a single forward pass.

Large-scale text-to-image (T2I) synthesis poses unique challenges, all of which are addressed by StyleGAN- T [19]. The specific requirements include extensive capacity, robust training across diverse datasets, precise text alignment and the ability to balance variation against text alignment according to user preferences. In terms of sample quality and speed, StyleGAN-T performs noticeably better than earlier GANs and surpasses distilled-diffusion models, which were the prior state-of-the-art models in quick T2I synthesis.

Diffusion models were first presented in [7], drawing inspiration from thermodynamics' non-equilibrium state. Fundamentally, in diffusion models, we gradually add Gaussian noise and then figure out how to reverse it.

Encoders play a pivotal role in both text and image encoding, providing a bridge between disparate data

modalities. The encoder's function is to transform complex information from textual and image inputs into a compressed, abstract representation conducive to further analysis or synthesis. In the context of text encoding, natural-language processing techniques are employed to distill semantic meaning, contextual nuances and sentiment from textual data. Simultaneously, in image encoding, visual features, patterns and spatial arrangements are captured and encoded to represent the essence of the visual content. It has been demonstrated that contrasting models, such as CLIP [20], are able to learn stable representations of images that capture both style and meaning to create images by using these representations. An effective technique for learning picture representation from natural-language supervision is Contrastive Language-Image Pre-training (CLIP), which trains both a text encoder and an image encoder simultaneously to anticipate the right pairings of a batch of (image, text) training samples. The target dataset's class names or descriptions are embedded by the learnt text encoder, which then uses this information to create a zero-shot linear classifier at test time.

Two methods were merged by A. Ramesh et al. [21] to solve the text-conditional image-creation problem. They initially trained the CLIP image encoder inverted using a diffusion decoder. The inverter exhibited non-determinism and had the ability to generate several pictures that corresponded to a particular embedding. Beyond T2I translation, the existence of an encoder and its near inverse (the decoder) enabled other possibilities. Encoding and decoding an input image yield semantically identical output images, just like in GAN inversion. By inverting the interpolations of the input images' image embeddings, this technique also made it possible to interpolate between them. But one important benefit of employing the CLIP latent space is that, unlike GAN latent space, which requires trial and error and intensive manual analysis to find these directions, one can semantically edit images by moving in the direction of any encoded text vector. Moreover, the processes of encoding and decoding images offer instruments for discerning the aspects of the image that CLIP acknowledges or ignores. The authors coupled the CLIP-image embedding decoder with an earlier model that produced CLIP image embeddings from a given text caption in order to create a comprehensive generative model of images.

A new degree of flexibility is provided by T2I, which allows users to direct the creative process using natural language. Customizing these models to match user-supplied visual conceptions is still a difficult issue, though. Combining several personalized concepts into a single image, keeping a modest model size and preserving high visual fidelity while permitting creative flexibility are just a few of the difficult problems that face T2I penalization. The Where Pathway and the What Pathway were utilized by Y. Tewel et al. [22] to enhance the user's control over what and where objects should be present in the final image. Using a text encoder, a text input is first turned to a sequence of word embeddings, which is subsequently changed into a sequence of encodings. Next, these encodings are projected *via* the W_k and W_v cross- attention matrices. K routes or W_k, were used to direct objects to their proper locations in the final image. On the other hand, W_v, sometimes called V routes, determined what should be included in the final image.

2.3.2 Motion Dynamics

Motion dynamics in the realm of video generation encompass the representation and understanding of temporal changes, spatial relationships and the flow of motion within a sequence of frames. This concept involves capturing the evolution and transitions of objects or scenes over time, ensuring that the generated videos exhibit realistic and coherent motion patterns. Key considerations include modeling how objects move in relation to each other, recognizing various actions or activities and representing the flow of motion with attention to factors, such as acceleration, deceleration and changes in direction. Effective motion-dynamics modeling also accounts for long-term dependencies, ensuring that the generated videos maintain consistency and contextually relevant temporal sequences. Now let's discuss some of the pivotal methodologies employed to attain motion dynamics.

- Carl Vondrick et al. [23] harnessed the wealth of unlabeled video data to develop a robust model centered around scene dynamics, emphasizing its applicability in both video recognition tasks, such as action classification and video generation tasks, like future prediction. A key contribution lies in the introduction of a generative adversarial network with a spatio-temporal convolutional architecture, strategically designed to disentangle foreground and background components within scenes.
- MoCoGAN [24] explicitly addresses the distinction between content and motion dynamics. It

employs a unique framework where a sequence of video frames is generated by mapping random vectors, with each vector comprising fixed content and a dynamic motion component modeled as a stochastic process. The innovation lies in its adversarial learning scheme, integrating video discriminators with images, to achieve unsupervised motion and content decomposition. The framework excels in generating videos with consistent content yet diverse motion and *vice versa*, showcasing its prowess in capturing and manipulating intricate motion dynamics within generated content.

Table 2. Text-to-image models and their features.

Study & Year	Algorithm	Dataset	Advantages	Limitations	Accuracy
[21], 2022	unCLIP (AR), unCLIP (diffusion)	• MS-COCO	• Complex, diverse & realistic images	• Not good at binding attributes	• unCLIP (AR prior) 10.63 FID Score • unCLIP (Diffusion prior) 10.39 FID score
[22], 2023	Gated Rank-1	• MS	• Less overfit • Better Pareto front	• Over-generalization • High amount of prompt engineering required when combining two or more concepts	• 2.18±0.02 (SEM)
[19], 2023	StyleGAN-T	• CC12m • CC • YFCC 100m (filtered) • Redcaps • LAION-aesthetic- 6+	• Better than DM at low resolution	• Less resolution • Struggles in producing images	• 13.90 FID score

- While existing methods struggled with entangling content tasks and motion in a sole-generator network, Ximeng Sun et al. proposed Two-Stream Variational Adversarial Network (TwoStream- VAN) [25] that adopts a two-stream model to disentangle these tasks. By progressively generating and fusing multiscaler motion alongside corresponding spatial content, the model excels in creating clear and consistent motion, resulting in photorealistic videos.
- Kangyeol Kim et al. [26] introduced an innovative method that entails learning distinct distributions for motion and appearance. Unlike previous methods that discretize motion dynamics, the proposed model utilizes neural ODE to capture the continuous nature of physical-body motion. The two-stage approach involves generation of a sequence of key points using a noise vector and synthesizing videos based on this sequence and an appearance noise vector. The model outperforms recent baselines quantitatively and showcases versatile functionalities, like motion-transfer among different datasets and dynamic frame-rate conversion, indicating promising applications for diverse video-generation scenarios.

2.3.3 Cross-frame Attention

It is imperative to verify that the video accurately depicts the event, that every frame is part of the same film and that there are no jumps in the video. To do this, it is essential to make sure that the frames generated after this are comparable. Cross-frame attention [27][9][28] approaches by various models were used to accomplish this. Table 3 presents the summary of work done in this area. A key idea in the fields of deep learning and artificial intelligence is attention [29]. Regarding neural networks and natural language processing, attention can be conceptualized as a technique that facilitates the model's ability

to concentrate on key information while processing data. Think of it like a spotlight that is focused on the most important portions of the input data. The mathematical definition denoting attention is represented in Equation (4):

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (4)$$

The attention function takes Q, K and V as inputs and its scaling factor is $\frac{1}{\sqrt{d_k}}$. While Q, K and V in self-attention are all part of the same sequence, in cross-attention, K and V are diffusion's conditioning parameters that control generation.

Fundamentally, attention enables the model to give distinct parts in a sequence various levels of importance. For example, when parsing a statement, some words or phrases could be more important than others to grasp the sentence's meaning. The model can adaptively weigh these words or phrases by highlighting the important ones and downplaying the less important ones according to attention mechanisms. When exploring self-attention, where the model evaluates the relationships between components inside the same sequence or cross-attention, which enables it to take into account interactions between elements from separate sequences, this idea becomes quite potent. Many natural language processing tasks now perform much better because to these methods, which also make them more accurate and context-aware. Essentially, attention functions as a kind of spotlight for neural networks, assisting them in identifying and concentrating on the most important data, ultimately producing increasingly complex and contextually-aware models.

A new issue in video creation arises from the addition of time as a dimension. The creation of context-aware videos presents whole new difficulties. In addition to having to pay attention to what is being done, there may be other things requiring focus that might be challenging to manage.

Using a U-Net architecture, J. Ho et al. [28] were able to create longer scenes by utilizing autoregressive extension and classifier-free guidance, which improved text-image linkages. Interleaved Spatial Super Resolution Model (SSR) and Temporal Super Resolution Model are employed by J. Ho et al. [10]. That enabled them to produce 128x128 videos at 24 frames per second, equivalent to 128x768 frames.

In order to introduce the time dimension into a two-dimensional (2D) conditional network, Uriel Singer et al. [9] employed spatiotemporal layers, specifically pseudo-3d convolutional layers and pseudo-3d attention layers. Consequently, cross-frame focus and consistent video were guaranteed. Inspired by separable convolutions, they inserted a 1D convolution after every 2D convolutional layer [30]. This improved temporal information fusion and made greater use of text-to-image (T2I).

Use of Large Language Models or LLMs is a very novel concept that H. Fei et al. [27] proposed. By employing existing LLMs for better simulation that is much closer to reality and produces better results, LLMs are used for scene simulation, making them more performant. Using Dysen (Dynamic Scene Manager)-VDM, a three-layer system combines scene imagination, event-to-DSG conversion and action planning. These layers provide more complicated scenarios by allowing several activities to occur simultaneously by using ChatGPT to interpret the action in a scene. When compared to Make-a-Video, it works noticeably better in longer, more intricate sequences.

2.3.4 Frame Interpolation

Frame Interpolation in text-to-video generation is a crucial step that enhances the temporal flow and smoothness of the generated video sequence. This process involves creating intermediate frames between existing frames to fill the gaps and improve the overall frame rate. By intelligently estimating the content and motion dynamics between consecutive frames, frame interpolation ensures a more fluid and natural transition in the video, contributing to a visually coherent and realistic output. This technique is particularly valuable in scenarios where the original frame rate is low or when generating high-quality videos to achieve a more lifelike and dynamic visual experience from the textual input.

1) Adaptive Separable Convolution:

- Niklaus et al. [30] employed a unique method to utilize a fully convolutional neural network to predict spatially adaptive convolutional kernels for each pixel, eliminating the need for independent-motion estimation and resampling stages. This approach efficiently captures local movement, rendering it resistant to occlusion, brightness fluctuations and blur.

Table 3. Algorithms for cross attention and their features.

Study & Year	Algorithm	Dataset	Advantages	Limitations	Accuracy
[27], 2022	Dysen (Dynamic Scene Manager)-VDM	• UCF-101 • MSR-VTT	• Performs better in scenarios with complex actions. • Uses LLMs for better dynamic	• Depends on ChatGPT	• IS 95.23 • FVD 255.42
[9], 2022	Spatiotemporal layers (pseudo- 3D convolution)	• WebVid-10M • HD-VILA-100M	• Better leverage a T2I architecture. • Allows for better temporal information fusion.	• Can not learn associations between text and phenomenon that can only be inferred in videos • Can generate short videos, and single scene/event.	• FVD 367 • IS 33
[28], 2023	U-Net classifier-free guidance autoregressive video extension	• Kinetics-600 • BAIR Robot Pushing • UCF101	• Enables joint training of text and video.	• Much worse compared to Make-A-Video in temporal information fusion	• FID 295±3 • IS 57±0.62
[10], 2022	SSR & TSR autoregressive video extension	• LAION-400M	• 128×128 videos at 24 frames per second equivalent to 128×768 frames.	• Lack of customizatio n options	• Clip Score 24.27 • Clip R-Precision 86.18

The introduction of separable convolutions significantly reduces processing requirements and the authors achieved substantial memory savings by estimating spatially adaptable 1D convolution kernels. This breakthrough improved previous methods, such as AdaConv, using a specialized encoder-decoder network to estimate kernels for all pixels simultaneously. The study addresses challenges in CNN-based frame interpolation, including occlusion management and resolution adaptation, marking a significant advancement in the field's effectiveness and accessibility.

- To mitigate computational complexity, Chen et al. [31] introduced deformable separable convolution (DSepConv). This technique aims to adaptively estimate kernels with suitable features to handle substantial motion. Subsequent enhancements in their model, known as EDSC [32], enabled the generation of numerous interpolated frames between consecutive frames. Nevertheless, achieving optimal performance for interpolating at arbitrary times remained challenging.

2) Path Selective Interpolation:

- Path Selective Interpolation is a powerful approach based on the principle that each pixel in interpolated frames follows a distinct path in preceding frames. Pioneered by Mahajan et al. [33], this method employs a path-based framework coupled with inverse optical flow

to calculate background motion. By moving and duplicating pixel gradients along anticipated paths, it minimizes issues like holes, chromatic aberrations and visual blur associated with traditional optical-flow techniques. Notably, path-based interpolation preserves the original frequency content and deterministically identifies veiled zones by prioritizing flow consistency, setting it apart from blending-based techniques.

- B. Yan et al. [34] incorporated standard optical-flow algorithms to the framework to control path direction and maintain global path coherency. They also introduced a pixel interlacing model to optimize optical-flow estimation, which significantly boosted efficiency.
- Y. Fan et al. [35] integrated semantic information acquired from input frames to identify crucial pixels *via* optimal-energy minimization, enhancing the precision of motion pattern detection. The inclusion of feature points from input frames resulted in more realistic and visually pleasing results. The approach, designed for processing larger input images and generating an arbitrary number of intermediate frames, aimed to provide exceptional visual quality.

3) Efficient Optical-flow Estimation:

- L. Khachatryan et al. [36] proposed an efficient optical-flow estimation method based on the local all-pass approach, operating in real time at high spatiotemporal resolutions. Using quadratic approximations, a higher-order approach compared to conventional first-order methods, the technique offers precise flow estimations. This unique methodology significantly enhances interpolated frame clarity by reducing motion boundary blur and preserving local geometric information. Notably, the approach addresses challenges in capturing fast, large- scale motion in optical flow-guided frame interpolation, employing a Laplacian cotangent mesh constraint for accurate motion representation, even in the presence of complex non- rigid motion. The implementation of a mesh system with one vertex per pixel demonstrates remarkable results in the Middlebury interpolation-error criterion, showcasing its potential applicability in optical flow-guided frame interpolation.

4) Real-time Frame Interpolation via GAN:

- J. van Amersfoort et al.'s work, "FIGAN" [37], represents a significant advancement in GAN- based frame interpolation. Demonstrating an impressive average runtime speedup of ×47 compared to rival approaches, FIGAN excelled in real-time YouTube 8M movies, establishing itself as the most sophisticated technique in the field. The authors introduced a multi-scale network with a mixed perceptual loss function, integrating spatial transformer networks with conventional optical-flow modeling. FIGAN's notable improvement over prior methods, such as SepConv-Lf, lies in its ability to generate higher-quality interpolated frames with fewer training parameters. This efficiency is particularly crucial in resource-limited scenarios, such as real-time video processing.
- S. Wen et al. [38] introduced a network comprising two concatenated GANs. The first GAN captures motion from training video clips and integrates finer frame data to enhance output quality, while the other generates frame details. To address issues related to noise that affected earlier approaches, they employed the Normalized Product Correlation Loss (NPCL). This innovative framework achieved visually appealing effects and demonstrated remarkable performance, particularly attributed to the effective use of NPCL, showcasing notable progress in the domain of GAN-based frame interpolation.
- J. Xiao et al.'s work, "FI MSAGAN"[39], used multi-scale dense attention generative adversarial networks to interpolate interim frames. FI MSAGAN accomplished more efficient fusion of local and global information details by using multiple generators and discriminator networks with varied sized input images. Its accuracy and runtime were found to be comparable to those of other state-of-the-art approaches.

5) Phase-based Frame Interpolation:

- P. Didyk et al. led the first study on phase-based frame interpolation. Their approach, as presented in [40], was based on the hypothesis that individual pixel phase-shift values

might contain limited motion information. However, the method struggled to effectively handle large movements, resulting in less than optimal results.

- S. Meyer et al. [41] developed a coarse-to-fine framework with a multi-scale pyramid level structure to communicate phase information. To address the issue of tolerating large motions, they capped phase-shift values. Phase-shift values were computed, phase differences were used to interpolate frames and amplitude values were used to blend the interpolated frames. Their algorithm was made up of these three essential steps. This method's inability to handle high- frequency motion resulted in blurry output in areas with small but high-frequency motion.
- The earliest phase-based techniques, including those introduced by the authors of [41], were characterized by the manual adjustment of parameters, such as amplitude and phase shift, to produce images. However, this manual adjustment process imposed limitations on the method's adaptability and efficiency. In order to estimate amplitude and phase-shift values directly, S. Meyer et al. [42] proposed the Phase-Net neural network architecture. Eliminating the need for manually adjusted parameters, this innovation significantly expanded the technique's ability to handle a broader spectrum of motion and frequencies. The authors used a decoder-only Phase-Net design, in which all levels were identical except for the last layer. Simulating a level-wise decomposition of phase information, the interpolated frame's resolution progressively grew as one proceeded up the network levels. By estimating parameters directly, Phase-Net was able to achieve more robust and diversified frame-interpolation capabilities than it could have achieved with hand-tuned phase-based approaches.

6) Bidirectional Optical Flow Estimation:

- H. E. Ahn et al. [43] proposed a method to effectively estimate bidirectional optical flow at lower resolutions and then reconstruct high-resolution optical flow. This multi-scale motion-reconstruction network works especially well with 4K footage and other high-resolution video frames. The method begins with bidirectional optical flow estimation at a lower resolution (e.g. one-fourth of the original resolution for 4K recordings). A multi-scale reconstruction strategy is then used to reconstruct the estimated optical-flow to match the original resolution. The authors trained their network using a variety of loss functions, such as adversarial loss, consistency loss and multi-scale smoothening loss. This all-encompassing strategy tack led the challenges of high-resolution video-frame interpolation and generated computationally efficient results while maintaining visual quality.
- In addition to interpolating frames, S. Y. Kim et al. [44] acknowledged the significance of boosting spatiotemporal resolution in contemporary videos. Both objectives were sought after by their combined model, which provided a thorough response to the requirements of high-resolution video footage.
- W. Bao et al. [45] used flow vectors and convolutional kernels to build an adaptive warping layer that generated output pixels by using both flow vectors and motion-compensation kernels. This method not only made frame interpolation better, but it also made other video enhancement methods, such as super-resolution, possible.
- By taking depth information into account, techniques such as depth-aware flow projection (DAIN) [46] specifically addressed issues with occlusion. Using depth-aware frame synthesis networks, kernel estimation and context extraction, DAIN presented an effective approach. With less parameters and more effective performance, DAIN produced impressive results by highlighting the significance of depth in frame interpolation.
- The incorporation of meta-learning approaches by M. Choi et al. [47], as well as the usage of attention networks by J. Xiao et al. [39] and M. Choi et al. [48], improved the efficiency of frame-interpolation techniques. These methods increased performance and efficiency by concentrating on attention and adaptation within feature representations.

3. DATASETS

Several datasets are instrumental in the development and evaluation of T2V generation models, each

offering unique challenges and characteristics that cater to specific research goals. Table 4 presents the summary of some popular datasets used in T2V generative models.

Table 4. Datasets used in T2V.

Dataset	Description
UCF-101 [49]	13,320 videos with 101 action categories; Realistic action videos collected form YouTube.
MSR-VTT [50]	10,000 videos with 20 classes; annotation of 20 sentences per video clip.
WebVid-10M [51]	10.7M video-caption pairs. Short videos with textual descriptions; 2.5 M video subset with a total of 52K video hours.
HD-VILA-100M [52]	100M video-caption pairs. 720p videos ranging over a total of 371.5K video hours.
Kinetics-600 [53]	480K video clips with 600 action classes; video duration of 10-sec.
BAIR Robot Pushing [54]	64x64 images of a robot pushing objects on a tabletop; conditioned on 2 frames, predicting 14 frames.

4. METRICS AND RESULTS DISCUSSION

Text-to-video generative models are commonly evaluated using metrics, such as Frechet Inception Distance (FID) [55], Clip score [20], among others. However, the Fréchet Video Distance (FVD) [4] metric stands out as a superior choice. FVD incorporates both visual quality and temporal coherence, providing a more comprehensive assessment of generated videos [1]. In contrast to FID, which only looks at static-picture quality, FVD takes into account the dynamic elements that are important for video assessment. As a result, FVD becomes a more reliable tool for evaluating the general coherence and integrity of produced video sequences. Thus, we employed it as the standard metric for our testing. FVD works on trajectories that reflect the routes of moving objects in the movies, drawing inspiration from the Fréchet distance used in curve-similarity evaluations. Through the application of the Fréchet distance concept to video analysis, FVD allows researchers to evaluate the entire motion patterns holistically and identify subtle changes or similarities across different video sequences. Understanding the underlying motion dynamics is essential for good video interpretation in a number of computer-vision disciplines, such as action recognition, anomaly detection and content-similarity evaluation, where its application is widespread.

Trajectory representation, spatial point correspondence and trajectory-distance evaluation are laborious steps in the computation of FVD. By using this technique, FVD provides a sophisticated assessment of video content, making it possible to spot minute changes in motion patterns that could go unnoticed by conventional video-comparison criteria. FVD is a useful tool for researchers and practitioners trying to quantify and comprehend the nuances of motion dynamics in video data; the lower the value, the more comparable the films. FVD is still a crucial indicator in the ever-evolving field of video analysis, helping progress areas like automatic video-content classification, human-behavior analysis and surveillance.

Instead of using Google's initial implementation [56], we adopted FVD, which was developed by StyleGAN- V [57]. It uses approximations for faster FVD calculation and the errors are within the range of 1e-6, which is a reasonal trade off. We used MSR-VTT [50] dataset which has 10000 videos and used the standard test-train split. For each test, the video shortest prompt (caption) was selected, so as to reduce test size and ensure faster testing and all test videos were scaled to 256x256 resolution. For all models in Table, 5 2990 videos with 16 frames each were generated. These were further scaled to 256x256 resolution to have uniformity in tests. All videos were then converted into frames and respective FVD scores for 16 frames were calculated.

For experimentation, we utilized an Nvidia RTX Quadro A5000 (24 GB VRAM), 64 GB RAM and an Intel Xenon 20 core CPU. Table 5 shows the performance (FVD scores) of the various pre-trained models along with their inference time. Show-1 [12] yielded the best results for 16 frames; however, it employed 4 models (1 generation model, 2 SR models, 1 interpolation model) and took the longest time (10min-12min) compared to other examined models (15sec-20sec) per video. Hotshot-XL [6] yielded

good results for 8 frames, but when tested for 16 frames, it performed significantly worse. We evaluated Text2Video- zero with three base T2I models and observed that better performing T2I models produced better T2V outcomes. The training steps of text-to-video models vary due to differences in architectures and methodologies. Show1 follows a modular approach, with distinct modules undergoing specific numbers of training steps: Keyframe Module (120,000 steps), Interpolation Module (40,000 steps) and First and Second Super-resolution modules (40,000 and 120,000 steps, respectively). Zero-shot models leverage pre-existing T2I architectures, eliminating the need for a training phase. Stable-Diffusion-v1.5, a base T2I model, underwent training over 595,000 steps at a resolution of 512x512. Dreamlike-diffusion-1.0 and Dreamlike- photoreal-2.0 are derived from Stable-Diffusion-v1.5, thereby inheriting its training characteristics. Potat1, a text-to-video finetuning model, undergoes around 2,500 steps with a consistent learning rate of 5e-6, leveraging ModelsScope's architecture for rapid adaptation.

Table 5. Comparison of various open-source models and their FVD scores along with their inference time.

Model	**FVD Score (fvd2048_16f)**	**Inference Time(sec/video)**
Text2Video-zero [5] (dreamlike-photoreal-2.0)	1420.9068	18.7556
Text2Video-zero [5] (dreamlike-diffusion-1.0)	1519.5902	18.2612
Text2Video-zero [5] (stable-diffusion-v1-5)	1498.2528	18.6525
Show-1 [12]	1094.6304	654.8715
Text-to-video-finetuning [15] (camenduru/potat1)	2132.1784	15.1471
Hotshot-XL [6]	1421.3931	19.9149

5. OPEN CHALLENGES

The field of T2V generation confronts several prominent challenges that impede the seamless transition from textual descriptions to visually coherent and compelling video content. One significant hurdle is the lack of coherence in the generated videos, which necessitates the development of advanced methods to ensure smooth transitions between frames and scenes, preventing abrupt changes and disjointed visual elements.

Penalization is another critical aspect that demands attention, with the need to explore techniques that can tailor generated videos to individual preferences and contextual details, making the content more engaging and relevant to diverse audiences. The persistent issue of low resolution in generated videos calls for innovative approaches to enhance visual quality, involving sophisticated upscaling methods and the preservation of fine details. Frame interpolation, a fundamental process in video generation, faces challenges related to the lack of intricate details, requiring solutions that produce smoother and more realistic transitions between frames. Background-smoothening techniques must be developed to eliminate artifacts and inconsistencies, ensuring a natural flow in the visual elements of the generated videos. Moreover, the field lacks comprehensive study and survey papers, which hinders a thorough understanding of existing research and limits the identification of critical gaps and opportunities for advancement.

Additionally, the language dependency on English presents a significant limitation, urging the exploration of models and approaches that can accommodate multiple languages to enhance inclusivity and accessibility. Addressing these multifaceted challenges collectively will propel the field towards the development of more coherent, personalized and culturally-aware text-to-video generation systems.

6. CONCLUSION AND FUTURE SCOPE

In conclusion, this article has provided a comprehensive overview of the advancements in text-to-video generation leveraging GANs and stable diffusion models. The synthesis of these two powerful techniques has demonstrated promising results in overcoming challenges associated with coherence, personalization and visual quality. GANs have proven effective in capturing intricate details and generating realistic video frames, while Stable Diffusion models contribute to stable and coherent video synthesis over extended sequences. The synergistic integration of these approaches holds great potential for addressing the limitations identified in the existing literature. As we move forward, it is imperative

to continue exploring innovative combinations of GANs and stable diffusion models, pushing the boundaries of text-to-video synthesis to new heights. Moreover, future-research directions should prioritize scalability, real-time processing and ethical considerations, ensuring the responsible development and deployment of these advanced techniques in diverse applications. The amalgamation of GANs and stable diffusion models signifies a promising trajectory for the evolution of text-to-video synthesis, offering a rich landscape of possibilities for researchers, practitioners and industries invested in multimedia content generation.

In the future, improved temporal modeling techniques within GANs and stable diffusion frameworks can be used to overcome the coherence challenge. Personalization gaps can be bridged by integrating attention mechanisms and reinforcement learning for a more nuanced understanding of individual preferences. To tackle low resolution, exploring novel upscaling methods, incorporating perceptual loss functions and fine-tuning architectures can significantly enhance visual quality. Future studies should focus on creating versatile models capable of handling diverse content types through domain adaptation and cross- modal learning. In summary, the future scope lies in integrating cutting-edge technologies to create more coherent, personalized and high-resolution text-to-video generation systems.

REFERENCES

[1] A. Singh, "A Survey of AI Text-to-Image and AI Text-to-Video Generators," arXiv preprint, arXiv: 2311.06329, Nov. 2023.
[2] Z. Xing et al., "A Survey on Video Diffusion Models," arXiv preprint, arXiv: 2310.10647, October 2023.
[3] I. J. Goodfellow et al., "Generative Adversarial Nets," Advances in Neural Information Processing Systems, arXiv: 1406.2661, pp. 2672–2680, 2014.
[4] T. Unterthiner et al., "Towards Accurate Generative Models of Video: A New Metric & Challenges," arXiv preprint, arXiv: 1812.01717, 2018.
[5] L. Khachatryan et al., "Text2Video-zero: Text-to-image Diffusion Models are Zero-shot Video Generators," arXiv preprint, arXiv: 2303.13439, March 2023.
[6] J. Mullan et al., "Hotshot-XL," [Online], Available: https://github.com/hotshotco/Hotshot-XL, October 2023.
[7] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan and S. Ganguli, "Deep Unsupervised Learning Using Nonequilibrium Thermodynamics," arXiv preprint, arXiv:1503.03585, 2015.
[8] Y. Song and S. Ermon, "Generative Modeling by Estimating Gradients of the Data Distribution," arXiv preprint, arXiv: 1907.05600, July 2019.
[9] U. Singer et al., "Make-a-video: Text-to-video Generation without Text-video Data," arXiv preprint, arXiv: 2209.14792, 2022.
[10] J. Ho et al., "Imagen Video: High Definition Video Generation with Diffusion Models," arXiv preprint, arXiv: 2210.02303, 2022.
[11] C. Saharia, "Photorealistic Text-to-image Diffusion Models with Deep Language Understanding," Proc. of the 36th Conf. on Neural Information Processing Systems (NeurIPS 2022), arXiv: 2205.11487, 2022.
[12] D. Junhao Zhang et al., "Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-video Generation," arXiv preprint, arXiv: 2309.15818, September 2023.
[13] Daquan Zhou et al., "MagicVideo: Efficient Video Generation with Latent Diffusion Models," arXiv preprint, arXiv: 2211.11018, November 2022.
[14] J. An et al., "Latent-Shift: Latent Diffusion with Temporal Shift for Efficient Text-to-video Generation," arXiv preprint, arXiv: 2304.08477, April 2023.
[15] J. Wang et al., "ModelScope Text-to-video Technical Report," arXiv preprint, arXiv: 2308.06571, August 2023.
[16] Z. Luo et al., "VideoFusion: Decomposed Diffusion Models for High-quality Video Generation," Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 10209-10218, 2023.
[17] S. Hong, J. Seo, S. Hong, H. Shin and S. Kim, "Large Language Models Are Frame-level Directors for Zero-shot Text-to-video Generation," arXiv preprint, arXiv: 2305.14330, May 2023.
[18] H. Huang, Y. Feng, C. Shi, L. Xu, J. Yu and S. Yang, "Free-Bloom: Zero-shot Text-to-video Generator with LLM Director and LDM Animator," arXiv preprint, arXiv: 2309.14494, September 2023.
[19] A. Sauer, T. Karras, S. Laine, A. Geiger and T. Aila, "Stylegan-t: Unlocking the Power of GANs for Fast Large-scale Text-to-image Synthesis," arXiv preprint, arXiv: 2301.09515, 2023.
[20] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras and Y. Choi, "Clipscore: A Reference-free Evaluation Metric for Image Captioning," arXiv preprint, arXiv: 2104.08718, 2022.
[21] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen, "Hierarchical Text-conditional Image Generation with Clip Latents," arXiv preprint, arXiv: 2204.06125, 2022.
[22] Y. Tewel, R. Gal, G. Chechik and Y. Atzmon, "Key-locked Rank One Editing for Text-to-image Personalization," arXiv preprint, arXiv: 2305.01644, 2023.

[23] C. Vondrick, H. Pirsiavash and A. Torralba, "Generating Videos with Scene Dynamics," arXiv preprint, arXiv: 1609.02612, 2016.
[24] S. Tulyakov, M.-Y. Liu, X. Yang and J. Kautz, "MoCoGan: Decomposing Motion and Content for Video Generation," Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, DOI: 10.1109/CVPR.2018.00165, Salt Lake City, USA, 6 2018.
[25] X. Sun, H. Xu and K Saenko, "TwoStreamVAN: Improving Motion Modeling in Video Generation," arXiv preprint, arXiv: 1812.01037, 2020.
[26] K. Kim et al., "Continuous-time Video Generation *via* Learning Motion Dynamics with Neural ODE," arXiv preprint, arXiv: 2112.10960, 2021.
[27] H. Fei, S. Wu, W. Ji, H. Zhang and T.-S. Chua, "Dysen-VDM: Empowering Dynamics-aware Text-to-video Diffusion with Large Language Models," arXiv preprint, arXiv: 2308.13812, 2023.
[28] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi and D. J. Fleet, "Video Diffusion Models," arXiv preprint, arXiv: 2204.03458, 2022.
[29] A. Vaswani et al., "Attention Is All You Need," arXiv preprint, arXiv: 1706.03762, 2017.
[30] S Niklaus, L Mai and F. Liu, "Video Frame Interpolation *via* Adaptive Convolution," arXiv preprint, arXiv: 1703.07514, 2017.
[31] X. Cheng and Z. Chen, "Video Frame Interpolation *via* Deformable Separable Convolution," Proc. of the AAAI Conf. on Artificial Intelligence, vol. 34, pp. 10607–10614, 2020.
[32] X. Cheng and Z. Chen, "Multiple Video Frame Interpolation via Enhanced Deformable Separable Convolution," IEEE Trans. on Pattern Analysis And Machine Intell., vol. 44, no. 10, pp. 7029-7045, 2021.
[33] D. Mahajan, F.-C. Huang, W. Matusik, R. Ramamoorthi and P. Belhumeur, "Moving Gradients: A Path-based Method for Plausible Image Interpolation," ACM Transactions on Graphics, vol. 28, no. 3, Article no.: 42, pp 1–11, 2009.
[34] B. Yan and Y. Chen, "Low Complexity Image Interpolation Method Based on Path Selection," Journal of Visual Communication and Image Representation, vol. 24, pp. 661–668, 2013.
[35] Y. Fan, N. Yoda, T. Igarashi and H. Ma, "Path-based Image Sequence Interpolation Guided by Feature Points," Proc. of the 2016 IEEE Int. Conf. on Image Processing (ICIP), DOI: 10.1109/ICIP.2016.7532421, Phoenix, USA, 2016.
[36] T. Jayashankar, P. Moulin, T. Blu and C. Gilliam, "Lap-based Video Frame Interpolation," Proc. of the 2019 IEEE International Conference on Image Processing (ICIP), DOI: 10.1109/ICIP.2019.8803484, Taipei, Taiwan, 2019.
[37] J. van Amersfoort et al., "Frame Interpolation with Multi-scale Deep Loss Functions and Generative Adversarial Networks," arXiv preprint, arXiv: 1711.06045, 2019.
[38] S. Wen et al., "Generating Realistic Videos from Keyframes with Concatenated GANs," IEEE Trans. on Circuits and Systems for Video Tech., vol. 29, pp. 2337–2348, 2019.
[39] J. Xiao and X. Bi, "Multi-scale Attention Generative Adversarial Networks for Video Frame Interpolation," IEEE Access, vol. 8, pp. 94842–94851, 2020.
[40] P. Didyk, P. Sitthi-Amorn, W. Freeman, F. Durand and W. Matusik, "Joint View Expansion and Filtering for Automultiscopic 3D Displays 3D Stereo Content Multiview Content," ACM Trans. on Graphics, vol. 32, no. 6, Article no. 221, pp. 1–8, 2013.
[41] S. Meyer, O. Wang, H. Zimmer, M. Grosse and A. Sorkine-Hornung, "Phase-based Frame Interpolation for Video," Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), DOI: 10.1109/CVPR.2015.7298747, Boston, USA, 2015.
[42] S. Meyer, A. Djelouah, B. McWilliams, A. Sorkine-Hornung, M. Gross and C. Schroers, "Phasenet for Video Frame Interpolation," arXiv preprint, arXiv: 1804.00884, 2018.
[43] H. E. Ahn, J. Jeong, J. Woo Kim, S. Kwon and J. Yoo, "A Fast 4K Video Frame Interpolation Using a Multi-scale Optical Flow Reconstruction Network," Symmetry, vol. 11, no. 10, Article no. 1251, 2019.
[44] S. Ye Kim, J. Oh and M. Kim, "FISR: Deep Joint Frame Interpolation and Super-resolution with a Multi-scale Temporal Loss," Proc. of the 34th AAAI Conf. on Artificial Intelligence (AAAI-20), pp. 11278-11286, 2019.
[45] W. Bao, W.-S. Lai, X. Zhang, Z. Gao and M.-H. Yang, "MEMC-Net: Motion Estimation and Motion Compensation Driven Neural Network for Video Interpolation and Enhancement," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 3, 2018.
[46] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao and M.-H. Yang, "Depth-aware Video Frame Interpolation," arXiv preprint, arXiv: 1904.00830, 2019.
[47] M. Choi, J. Choi, S. Baik, T. H. Kim and K. Mu Lee, "Scene-adaptive Video Frame Interpolation *via* Meta-learning," arXiv preprint, arXiv: 2004.00779, pp.9444-9453, 2020.
[48] M. Choi, H. Kim, B. Han, N. Xu and K. Mu Lee, "Channel Attention Is All You Need for Video Frame Interpolation," Proc. of the 34th AAAI Conf. on Artificial Intelligence (AAAI-20), pp. 10663- 10671, 2020.
[49] K. Soomro, A. R. Zamir and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild," Proc. of the 1st Int. Workshop on Action Recognition with Large Number of Classes, arXiv: 1212.0402, 2012.

[50] J. Xu, T. Mei, T. Yao and Y. Rui, "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), DOI: 10.1109/CVPR.2016.571, Las Vegas, USA, 2016.

[51] M. Bain, A. Nagrani, G. Varol and A. Zisserman, "Frozen in Time: A Joint Video and Image Encoder for End-to-end Retrieval," arXiv preprint, arXiv: 2104.00650, 2021.

[52] H. Xue et al., "Advancing High-resolution Video-language Representation with Large-scale Video Transcriptions," Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), DOI: 10.1109/CVPR52688.2022.00498, pp. 5026-5035, 2021.

[53] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier and A. Zisserman, "A Short Note about Kinetics-600," arXiv preprint, arXiv: 1808.01340, 2018.

[54] F. Ebert, C. Finn, A. X. Lee and S. Levine, "Self-supervised Visual Planning with Temporal Skip Connections," Proc. of the 1st Conf. on Robot Learning (CoRL 2017), Mountain View, USA, pp. 1-13, 2017.

[55] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter, "Gans Trained by a Two Time-scale Update Rule Converge to a Local Nash Equilibrium," Proc. of the 31st Conf. on Neural Information Processing Systems (NIPS 2017), pp. 1-12, Long Beach, USA, 2018.

[56] Github, "Google-research," [Online], Available: https://github.com/google-research/google-research/blob/master/frechet_video_distance.

[57] I. Skorokhodov, S. Tulyakov and M. Elhoseiny, "StyleGAN-V: A Continuous Video Generator with the Price, Image Quality and Perks of StyleGAN2," Proc. of the IEEE CVPR 2022, pp. 3626-3636, 2021.

ملخص البحث:

يتطلّب التّحدي المتمثّل في السّعي إلى توليد الفيديو من النّصوص تحويل الأوصاف النّصّية إلى مقاطع فيديو حقيقية ومتماسكة. ولقد تعّرض هذا المجال البحثي إلى تطوّرات عديدة في السنوات الأخيرة، مع تطوّر النّماذج الاندماجية والشّبكات الاستدراكية التّوليدية (GANs).

تبحث هذه الورقة في أحدث نماذج التّحويل من نُصوص إلى صور فيديو، والخطوات الّتي يتضمّنها توليد صور الفيديو من النّصوص، بما فيها التّماسُك المؤقت، وتوليد الفيديو، إلى جانب ترميز النّصوص. كذلك نتناول التّحدّيات الّتي ينطوي عليها تحويل النّصوص إلى فيديو وأحدث ما تمّ التّوصّل إليه من الطّرق للتغلّب عليها. هذا إلى جانب تحليل ومراجعة مجموعات البيانات والمقاييس التي يكثر استخدامها في هذا المجال.