ISSN No:-2456-2165

Automatic Video Generator

K Tresha[1]; Kavya[2]; Medhaa PB[3]; Pragathi T[4]
Department of Information Science RNS Institute of Technology, Bengaluru

Abstract:- Text-to-video (T2V) generation is an emerging field in artificial intelligence, gaining traction with advances in deep learning models like generative adversarial networks (GANs), diffusion models, and hybrid architectures. This paper provides a comprehensive survey of recent T2V methodologies, exploring models such as GAN-based frameworks, VEGAN-CLIP, IRC-GAN, Sora OpenAI, and CogVideoX, which aim to transform textual descriptions into coherent video content. These models face challenges in maintaining semantic coherence, temporal consistency, and realistic motion across generated frames. We examine the architectural designs, methodologies, and applications of key models, highlighting the advantages and limitations in their approaches to video synthesis. Additionally, we discuss benchmark advancements, such as T2VBench, which plays a crucial role in evaluating temporal consistency and content alignment. This review sheds light on the strengths and limitations of existing approaches and outlines ethical considerations and future directions for T2V generation in the realm of generative AI.

Keywords:- *Text-to-Video (T2V) Generation, Deep Learning, Generative Adversarial Networks (GANs), Diffusion Models, Hybrid Architectures, VQGAN-CLIP,IRC-GAN, Sora Open AI, Cog Video X, Semantic Coherence, Temporal Consistency, Realistic Motion, Video Synthesis, Benchmark Advancements, T2VBench,Content Alignment, Ethical Considerations, Generative AI.*

I. INTRODUCTION

Text-to-video (T2V) generation represents a cutting-edge area in multimedia content creation, where generative models aim to translate textual descriptions into dynamic, visually coherent videos. Unlike static image generation, video synthesis involves not only creating realistic visuals but also ensuring temporal coherence across frames, which adds considerable complexity to the task. Recent advances in generative models, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models, have enabled significant progress in generating high-quality, realistic videos from textual prompts.

The rise of deep learning has transformed content creation across media, positioning generative AI as a powerful tool for video synthesis. This task is challenging as it requires models to capture both spatial and temporal features, translating a single textual prompt into a sequence of visually consistent and contextually accurate frames. Models like VQGAN-CLIP, Temporal GANs Conditioning on Captions (TGANs-C), and hybrid VAE-GAN architectures have been designed to address these unique challenges, focusing on semantic coherence, smooth transitions, and realistic motion.

Despite these advancements, T2V generation faces several limitations. Models often struggle with maintaining high resolution, stable frame quality, and semantic alignment with input captions across sequences. High computational costs further constrain the scalability and accessibility of these methods. Nonetheless, new models such as IRC-GAN, Sora OpenAI, and CogVideoX have made strides in improving video coherence and alignment with input prompts. Benchmarks like T2VBench have emerged to provide a standard for evaluating temporal consistency and content alignment, both crucial for producing realistic video content.

This paper surveys recent methodologies in T2V generation, analyzing their architectures, strengths, and limitations. By exploring advancements in T2V models and the emerging trends in this field, we aim to highlight the current capabilities of generative AI in video synthesis and identify areas for future improvement.

II. LITERATURE SURVEY

➢ *Introduction to Text-to-Video Generation:*

Text-to-video (T2V) generation is an evolving field that translates textual descriptions into video outputs, bridging the gap between linguistic and visual information. Advances in generative AI, especially in generative adversarial networks (GANs) and diffusion models, have been instrumental in addressing the challenges of maintaining temporal coherence and scene consistency across frames.

➢ *Generative Models for T2V Synthesis:*

The main generative models facilitating T2V synthesis include GAN-based architectures, vector quantized GANs combined with CLIP (VEGAN-CLIP), and diffusion models. Each model addresses the unique demands of video generation, such as maintaining spatial features across frames and ensuring temporal continuity. Notable approaches include:

- TiVGAN: Generates frames sequentially from a single image, focusing on evolving visual coherence through an iterative process.
- TGANs-C: Introduces multi-discriminator GANs for text-aligned temporal coherence.
- Tune-A-Video: Leverages pre-trained text-to-image models fine-tuned with single text-video pairs for efficiency.

➢ *Evaluation and Benchmarking:*

To ensure quality in T2V generation, benchmark systems like T2VBench evaluate models on temporal consistency, alignment accuracy, and narrative flow. T2VBench assesses models such as ZeroScope and Pika on 16 performance dimensions, helping highlight each model's specific strengths, such as sequencing and movement dynamics.

➢ *Key Challenges and Future Directions:*

Despite promising advancements, challenges such as maintaining semantic coherence and producing high-resolution outputs persist. Future research directions include incorporating multimodal data, such as audio, and enhancing spatio-temporal consistency to improve narrative flow and fidelity in generated videos.

This overview integrates methodologies, benchmarks, and future insights based on recent research developments in text-driven video generation.

➢ *Objectives*

- *Develop Advanced Generative Models:*

To create robust generative models, particularly leveraging GANs, VAEs, and diffusion models, capable of translating textual descriptions into realistic and coherent video sequences. The focus is on achieving high-quality synthesis that respects temporal coherence across frames and aligns accurately with input text.

- *Enhance Temporal and Spatial Coherence:*

To improve the temporal consistency and spatial fidelity of generated videos, ensuring smooth transitions between frames and realistic motion. Techniques such as multi-discriminator frameworks, spatio-temporal attention, and structure-guided sampling are explored to maintain object and motion continuity across sequences.

- *Benchmark and Evaluate T2V Models:*

To establish comprehensive evaluation benchmarks, like T2VBench, for assessing T2V models across multiple dimensions including event sequencing, narrative flow, and alignment accuracy. This enables a structured comparison of models and identification of areas needing improvement.

- *Reduce Computational Requirements:*

To optimize the efficiency of T2V models by developing methods that reduce computational costs, such as one-shot tuning and efficient fine-tuning of pre-trained text-to-image models. This is crucial to make T2V generation accessible and practical for broader applications.

- *Broaden Application Scope and Real-World Relevance:*

To explore applications beyond synthetic datasets, such as real-world scenarios that demand high interactivity, semantic understanding, and multimodal data integration. This includes adapting T2V models for use in entertainment, education, and personalized content creation.

➢ *Proposed System*

The proposed systems from the documents cover various advanced generative models for text-to-video synthesis. Here's a summary:

- TiVGAN (Text-to-Image-to-Video Generative Adversarial Network) - TiVGAN generates videos through an evolutionary process, creating an initial high-quality frame from a text input and sequentially evolving it to ensure coherence across frames. Its staged training stabilizes frame quality but can struggle with complex, dynamic scenes.

- TGANs-C (Temporal GANs Conditioning on Captions)- TGANs-C uses a multi-discriminator GAN framework to ensure temporal and semantic coherence with text captions. It leverages video, frame, and motion discriminators to maintain smooth transitions and realistic content but has high computational demands.

- Tune-A-Video - This model extends pretrained text-to-image models for video generation by finetuning them on a single text-video pair, making it computationally efficient. However, its one-shot tuning approach limits generalization without additional data.

- IRC-GAN (Introspective Recurrent Convolutional GAN) - This model combines a recurrent generator with LSTM and convolutional layers for better alignment with text and is suited for high-resolution tasks but lacks scalability for real-time applications.

- Sora OpenAI - Designed for democratized video creation, Sora uses a transformer-based diffusion model, enabling high-resolution, complex scene generation. Ethical concerns around misuse highlight the need for safeguards.

- CogVideoX - Featuring a 3D VAE and a diffusion transformer, CogVideoX generates long, coherent videos with high dynamic object realism but is limited by computational demands.

These models demonstrate progress in aligning video content with text inputs while balancing quality and efficiency. Challenges remain, particularly with temporal coherence, resolution, and ethical use.

➢ *Advantages of Proposed System*

The proposed systems for text-to-video generation offer several advantages, each designed to enhance the quality, coherence, and efficiency of generated videos:

- *TiVGAN (Text-to-Image-to-Video Generative Adversarial Network)*

✓ Enhanced Temporal Coherence: By generating each frame sequentially, TiVGAN improves coherence across frames, maintaining smooth transitions.

ISSN No:-2456-2165

✓ Stability in Training: The stepwise frame-generation approach helps stabilize training, reducing issues like mode collapse common in GANs.

- *TGANs-C (Temporal GANs Conditioning on Captions)*

✓ Robust Semantic and Temporal Alignment: TGANs-C uses a multi-discriminator setup to ensure that each frame aligns with the text input while maintaining video continuity.
✓ Smooth Transitions: Its use of motion and video discriminators allows for more realistic transitions, improving the fluidity of generated motion.

- *Tune-A-Video*

✓ High Efficiency: Tune-A-Video is highly efficient, leveraging pretrained text-to-image models and requiring only one-shot tuning, which reduces computational costs significantly.
✓ Object Consistency Across Frames: By focusing on spatio-temporal attention, it maintains object consistency, resulting in more stable and coherent videos.

- *IRC-GAN (Introspective Recurrent Convolutional GAN)*

✓ High-Resolution Output: IRC-GAN's architecture supports high-resolution output, making it suitable for applications demanding quality detail.
✓ Improved Frame Alignment with Text: The introspective mechanism ensures that frames are closely aligned with textual inputs, which is crucial for applications needing precise semantic fidelity.

- *Sora OpenAI*

✓ Complex Scene Handling: Sora's transformer-based diffusion model enables it to handle complex, minute-long videos with consistent frame quality, suitable for industries like entertainment and education.
✓ Democratized Access to Video Creation: Designed with accessibility in mind, it allows more users to create high-quality videos from text prompts, making it useful for diverse fields.

- *CogVideoX*

✓ Long-Form, Coherent Video Sequences: CogVideoX excels at generating long sequences with dynamic object tracking and scene realism, ideal for creating extended content.
✓ Advanced Resolution and Detail: With multi-resolution frame packing and progressive training, it produces detailed, high-quality video output.

These models push the boundaries of text-to-video generation by enhancing the realism, coherence, and efficiency of video synthesis, making them suitable for applications in entertainment, marketing, education, and beyond.

III. PROPOSED METHODOLOGY

The proposed methodology for these advanced generative models combines several novel and creative approaches to tackle the complex challenges of text-to-video generation. Here's an outline of the main methodologies used by each model:

➢ *TiVGAN (Text-to-Image-to-Video Generative Adversarial Network)*

- Sequential Frame Generation: TiVGAN initiates the video by generating a high-quality single frame based on text input, then incrementally adds frames. This evolutionary approach allows the model to stabilize its output progressively, focusing first on visual accuracy and later on achieving temporal consistency across frames.

➢ *TGANs-C (Temporal GANs Conditioning on Captions)*

- Multi-Discriminator Architecture: TGANs-C employs a video, frame, and motion discriminator to enforce both spatial and temporal coherence in the generated video. These discriminators work together to assess frame-by-frame realism and the overall sequence continuity, ensuring the video aligns with text input in a seamless, continuous flow.

➢ *Tune-A-Vide*

- One-Shot Tuning with Pretrained Text-to-Image Models: Tune-A-Video reuses pretrained text-to-image models, tuning them with a single text-video pair. It employs a diffusion-based sampling method with spatio-temporal attention, which is a lightweight way to expand text-to-image capabilities into video, maintaining consistency while using minimal data and computation.

➢ *IRC-GAN (Introspective Recurrent Convolutional GAN)*

- Recurrent Convolutional and LSTM Layers: By integrating recurrent layers (LSTM) with 2D convolutions, IRC-GAN enhances frame quality and enforces temporal coherence. Its introspective mechanism maximizes mutual information between frames, aligning them effectively with textual prompts. This makes it particularly effective for high-resolution tasks where detail and frame alignment are crucial.

➢ *Sora OpenAI*

- Transformer-Based Diffusion Model: Sora uses a transformer model to structure high-quality, complex, minute-long videos that retain frame consistency and detail. By incorporating feedback loops for continuous refinement based on user input, it adapts to different video types, making it versatile for various applications, like marketing and education.

➢ *CogVideoX*

- Diffusion Transformer with 3D VAE: CogVideoX uses a 3D Variational Autoencoder (VAE) in combination with a diffusion transformer model. This method allows for long-form generation with high realism, using a multi-resolution frame-packing approach to produce detailed, coherent scenes while progressively training the video sequence for quality enhancement.

Together, these methodologies showcase creative uses of GANs, transformers, VAEs, and one-shot tuning to generate videos that are semantically aligned, temporally consistent, and computationally efficient. The blend of stepwise generation, multi-discriminator setups, and progressive training strategies marks a breakthrough in AI-driven video generation, opening up vast new applications and possibilities.

➢ *System Architecture*

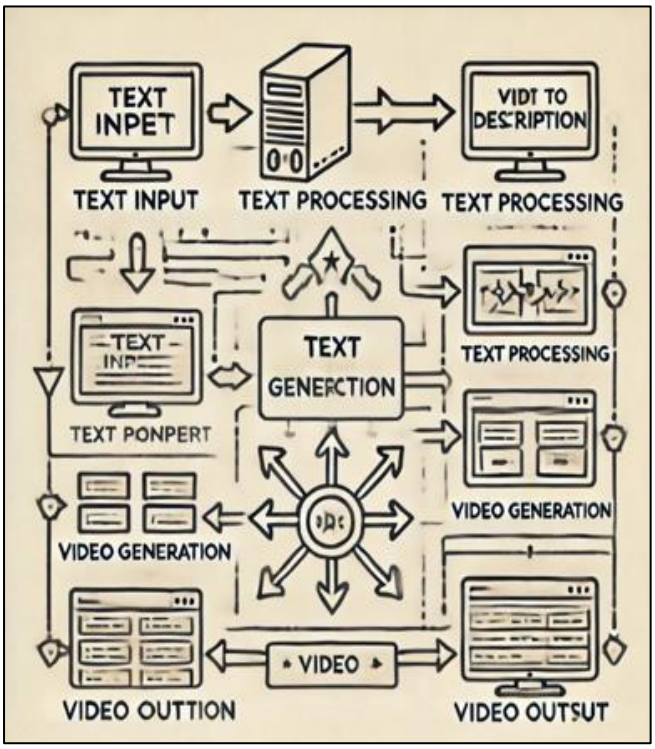

Fig 1 System Architecture

IV. CONCLUSION

Models like TiVGAN, TGANs-C, Tune-A-Video, IRC-GAN, Sora OpenAI, and CogVideoX have addressed challenges in generating realistic, coherent videos from text, each contributing unique solutions to improve temporal coherence, semantic alignment, and resolution.

While these models exhibit significant strengths—such as efficient training, semantic consistency, and handling complex scenes—limitations still exist. High computational costs and challenges in maintaining long-term coherence and quality, especially in dynamic scenes and high resolutions, remain hurdles to broader adoption.

REFERENCES

[1]. TiVGAN: Text to Image to Video Generation With Step-by-Step Evolutionary Generator DOYEON KIM(Member, IEEE), DONGGYU JOO AND JUNMO KIM , (Member, IEEE)School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea.

[2]. Generate Impressive Videos with Text Instructions: A Review of OpenAI Sora, Stable Diffusion, Lumiere and Comparable Models by Enis Karaarslan1 and ¨Omer Aydın1.

[3]. Conditional GAN with Discriminative Filter Generation for Text-to-Video.Synthesis by Yogesh Balaji ,Martin Renqiang Min , Bing Bai , Rama Chellappa1 and Hans Peter Graf2.University of Maryland, College Park, NEC Labs America – Princeton

[4]. Transforming Text into Video: A Proposed Methodology for Video Production Using the VQGAN-CLIP Image Generative AI Model by SukChang Lee *Prof., Dept. of Digital Contents, Konyang Univ., Korea*

[5]. To Create What You Tell: Generating Videos from Captions by Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li and Tao Mei.University of Science and Technology of China, Hefei, China.Microsoft Research, Beijing, China

[6]. Yitong Li, Martin Renqiang Min,Dinghan Shen, David Carlson,Lawrence Carin,Duke University, Durham, NC, United States, 27708 NEC Laboratories America, Princeton, NJ, United States, 08540 {yitong.li, dinghan.shen, david.carlson, lcarin}@duke.edu, renqiang@nec-labs.com

[7]. AUTOLV:AUTOMATIC LECTURE VIDEO GENERATOR *Wenbin Wang Yang Song Sanjay Jha ,School of Computer Science and Engineering, University of New South Wales, Australia*

[8]. Sounding Video Generator: A Unified Framework for Text-guided Sounding Video Generation.Jiawei Liu, Weining Wang, Sihan Chen, Xinxin Zhu, Jing Liu

[9]. IRC-GAN: Introspective Recurrent Convolutional GAN for Text-to-video Generation,Kangle Deng , Tianyi Fei, Xin Huang and Yuxin Pengy.Institute of Computer Science and Technology, Peking,University, Beijing, China.pengyuxin@pku.edu.cn

[10]. Sora OpenAI's Prelude: Social Media Perspectives on Sora OpenAI and the Future of AI Video Generation:REZA HADI MOGAVI, DERRICK WANG, JOSEPH TU, HILDA HADAN, and SABRINA A.

[11]. SGANDURRA,Stratford School of Interaction Design and Business, University of Waterloo, Canada,PAN HUI, Hong Kong University of Science and Technology (Guangzhou), Hong Kong SAR and Guangzhou, China,LENNART E. NACKE, Stratford School of Interaction Design and Business, University of Waterloo, Canada
[12]. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer: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 Jie Tang
[13]. StreamingT2V: Consistent, Dynamic, and Extendable.Long Video Generation from Text:Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan,Zhangyang Wang1,2, Shant Navasardyan1, Humphrey Shi1,3,1Picsart AI Research (PAIR) 2UT Austin 3SHI Labs @ Georgia Tech, Oregon & UIUC
[14]. TAVGBench: Benchmarking Text to Audible-Video Generation:Yuxin Mao1, Xuyang Shen2, Jing Zhang3, Zhen Qin4, Jinxing Zhou5, Mochu Xiang1, Yiran Zhong2, Yuchao Dai1.Northwestern Polytechnical University
[15]. ,OpenNLPLab, Shanghai AI Lab ,Australian National University,TapTap 5Hefei University of Technology ART•V: Auto-Regressive Text-to-Video Generation with Diffusion Models:Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin,Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, Chong Luo, Yueyi Zhang, Zhiwei Xiong.University of Science and Technology of China ,Microsoft Research Asia
[16]. Rescribe: Authoring and Automatically ,Editing Audio Descriptions**:**Amy Pavel ,Gabriel Reyes ,Jeffrey P. Bigham T2VBench: Benchmarking Temporal Dynamics for Text-to-Video Generation by Pengliang Ji, Chuyang Xiao, Huilin Tai, Mingxiao Huo.Carnegie Mellon University,ShanghaiTech University,McGill University
[17]. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation by Jay Zhangjie Wu Yixiao Ge Xintao Wang Stan Weixian Lei Yuchao Gu Yufei Shi Wynne Hsu Ying Shan Xiaohu Qie Mike Zheng Shou.Show Lab, National University of Singapore ARC Lab, Tencent PCG
[18]. LAVIE: HIGH-QUALITY VIDEO GENERATION WITH CASCADED LATENT DIFFUSION MODELSYaohuiWang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang,Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang,Yuwei Guo, TianxingWu, Chenyang Si, Yuming Jiang, Cunjian Chen,Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, Ziwei Liu
[19]. CogVideo: Large-scale Pre Training for Text-to-Video,Generation via Transformers by Wenyi Hong,Ming Ding,Wendi Zheng, Xinghan Liu, Jie Tang,Tsinghua University zBAAI {hongwy18@mails, dm18@mails, jietang@mail}.tsinghua.edu.cn
[20]. To Create What You Tell: Generating Videos from Captions by Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li and Tao Mei.University of Science and Technology of China, Hefei, China.Microsoft Research, Beijing, China