Text powered video generation using stability AI

Benakappa S M[1], Moulya R G[2], Poorvika[2], Pratheeksha D R[2], Rakshitha R[2]
[1]Associate Professor, Department of Computer Science and Engineering, JNNCE Shivamogga, Karnataka, India
[2]UG Students, Department of Computer Science and Engineering, JNNCE, Shivamogga, Karnataka, India

***Abstract*— Turning text into rich multimedia content using the latest Generative AI (GenAI) technologies is the utmost requirement in today's world. It is designed to meet the growing demand for automated content creation in areas such as digital media, education, marketing, and entertainment. The system works by taking a user's text description and generating relevant images. These images are then stitched together to create a smooth, meaningful video. By using advanced AI models, the process becomes easier and more efficient, helping creators bring their ideas to life with less effort. The paper reveals an extensive survey conducted on cutting-edge techniques like latent diffusion models for generating images, transformer-based models to understand the text, and specialized tools to create videos that are both visually consistent and fluid.**

Further, a survey aids to devise a model in delivering high-quality results that stay true to the original input.

Index Terms– *Stable Diffusion, Stability AI, GEN AI, Large Language Model, Latent Diffusion Model*

I.INTRODUCTION

The semantics of the text is the first step in transforming a straightforward text prompt into a meaningful image. Using potent pre-trained language-vision models such as Contrastive Language–Image Pre-training or T5, natural language input is transformed into a rich, high-dimensional representation at its inception. The intent and meaning of the words are better conveyed by these models. Latent diffusion models (LDMs), which are renowned for efficiently generating high-quality images by operating in a compressed, latent space, are then guided by this semantic embedding. This method enables precise control over the images' style and content, enabling tasks like scene editing and artistic transformation to be completed with remarkable accuracy [1]. The text-to-image synthesis is now much more flexible and accurate thanks to recent advancements in generative modeling. It is noteworthy that multimodal conditioning methods that combine hand-drawn sketches and textual descriptions have produced outputs with better visual realism and structural accuracy [2]. At the same time, user-driven customization has become a crucial component, allowing for fine-grained control over characteristics like age, emotion, and hairstyle, thus increasing the model's applicability in a variety of fields such as fashion, medical [3]. Simultaneously, a number of studies have evaluated the shortcomings of current text-to-image systems critically, highlighting the need for better contextual understanding and the availability of large-scale, high-quality training datasets [4]. The recent methods have addressed these issues by introducing context-aware generation frameworks that improve output consistency and semantic coherence, especially when working with complex or abstract textual prompt [5]. The above methods collectively form the foundation of a pipeline for creating images. They demonstrate how far the field has advanced in converting abstract descriptions into striking, meaningful images, thereby bridging the gap between language and vision.

Further, these static keyframes are converted into continuous motion in order to create coherent video sequences, which are based on the synthesis of high-quality images from text prompts. Either a series of text-guided keyframes or temporally dynamic text-to-video diffusion models can be used to accomplish the task. A multimodal approach that uses both textual input and motion structure elements like pose or semantic segmentation to direct the video generation process has been introduced in order to preserve temporal consistency and motion alignment [6]. Additionally, architectural improvements have been suggested, such as frame consistency modules and spatiotemporal attention mechanisms, which greatly increase the overall stability and quality of produced videos by maintaining object integrity across frames [7]. The Large Language Models (LLMs) are incorporated to enhance semantic content and offer contextual cues throughout the sequence, and a U-Net-based denoising pipeline, common in diffusion models, has also been used to minimize visual artifacts and smooth transitions [8]. A crucial method for transforming sparse image sets into realistic and flowing video streams, intermediate frame prediction allows for more seamless transitions and consistent motion between keyframes [9]. When taken as a whole, these developments enable text-to-video generation that is both visually consistent and narratively aligned, providing increased control and aesthetic fidelity in the final products.

The proposed system's fundamental approach combines several generative frameworks into a modular, end-to-end pipeline that transforms text prompts into logical video outputs. In order to extract rich semantic embeddings, user-provided text inputs are initially processed using pretrained language encoders such as CLIP or T5. The latent diffusion models produce high-fidelity images in line with the text, are conditioned by these embeddings. Each step refines motion and temporal coherence by sequentially transforming static images into dynamic video frames, as shown in a previous stepwise video generation approach [8]. Following the creation of the initial image frames, smooth transitions are produced by predicting intermediate frames using video interpolation

techniques, specifically a hierarchical refinement method for multi-frame generation [3]. Additionally, a new context-enhanced video synthesis model [9] combines U-Net-based denoising with LLMs to improve temporal alignment and semantic consistency. The proposed system also includes flexible components based on stable diffusion mechanisms [10] to facilitate scene-level control and personalization, allowing for fine-grained adjustments like altering ambient features or facial expressions. From textual encoding to final video rendering, this tiered approach guarantees precise alignment between each step while preserving adaptability and scalability.

One important source for Stability AI's open-source generative models, which include popular systems like Stable Diffusion, Stable Video Diffusion (SVD), and, is the Stability-AI/generative-models repository on GitHub. The repository identifies a number of persistent technical and operational issues, even though these models mark significant breakthroughs in text-to-image and text-to-video generation. Among the most prominent are persistent memory management problems, which often appear as "Compute Unified Device Architecture *out of memory*" errors even on top-tier GPUs such as the A100 and RTX 3090. These issues indicate inefficiencies in memory allocation and training processes [11]. Simultaneously, the absence of thorough documentation, particularly for parameters like motion bucket_id, reflects more general usability issues that are mirrored [6], where the incorporation of structural and textual guidance enhances user control and accessibility. The model interaction is through retrieval-augmented generative mechanisms [12], the setup and configuration complexity are further increased by disjointed guidance on hardware compatibility and training pipelines. Beyond technical constraints, the unintentional inclusion of explicit content in training datasets has given rise to ethical questions regarding data provenance and curation standards [13]. The issues highlight the necessity of improved documentation, memory optimization, and legally binding ethical standards to guarantee responsible innovation in open-source generative AI.

A variety of technical and usability-focused solutions have been implemented to address the main issues noted in the Stability AI Generative Models repository. In order to reduce Graphic Processing Unit load during training and inference, memory optimization techniques like attention-slicing, activation checkpointing, and architectural improvements have been used. These techniques are similar to those investigated in [11] for effective high-resolution video generation. Better inline code comments, more thorough documentation, and improved configuration guides have all improved usability. This is especially helpful in helping users understand complicated parameters like motion_bucket_id, a problem that Make-Your-Video [6] also addresses by emphasizing user-controllable and interpretable interfaces. Iterative improvement and real-time troubleshooting are supported by community engagement through pull requests, GitHub issues, and FAQs. Dataset filtering and greater transparency help to mitigate ethical concerns, particularly with regard to the unintentional creation of NSFW content. This is consistent with best practices emphasized [13], where ethical dataset construction is prioritized. Additionally, the diffusers library and integration with the Hugging Face ecosystem have greatly simplified deployment processes and accessibility, and retrieval-augmented mechanisms suggested [12] encourage enhancements in content safety and relevancy. Together, these advancements seek to establish a framework for generative AI research and application that is more scalable, effective, and morally sound.

II. LITERATURE SURVEY

The survey of literature looks at progress in Generative AI for multimedia content. It calls out techniques augmenting text-to-video creation by LLM direction and latent motion flow, and progress in text-to-image diffusion models for better control and efficiency. The overall body of research seeks to improve fidelity, semantic accuracy, and personalization in AI-generated content.

In [14], an extensive survey on the history and mechanisms of text-to-image diffusion models. They gathered a thorough analysis of different control methods being used in these models, which helps researchers grasp how to increase control over output images. This study is a useful tool for those seeking to enhance the accuracy and consistency of the generated images. Although the results yield important findings, the research is more theoretical and does not have novel experimental verifications, which could decrease its real-world usability.

In [15], the authors proposed latent diffusion methods for image synthesis by working within a compressed latent space. Their approach offers a computationally effective alternative to classic pixel-space diffusion approaches, while enabling the preservation of high-resolution outputs with computational efficiency. Its salient contribution lies in the capability of balancing between quality and performance, hence catering to a wide range of applications. A pointed shortcoming, however, is that it may not completely preserve very fine-grained features, which might prove crucial to some high-fidelity image synthesis tasks.

The authors [16] proposed a model that creates videos from static images with latent motion flow guidance. The new method facilitates conditional video generation by learning the possible movement of pixels in a given image and therefore creates videos with coherent motion and appearance. The value of their work is the fact that it can create dynamic content with the desired motion, improving the storytelling value of video generation. However, the model is not very good at showing large or random movements, something that may restrict its use in more complicated situations.

In [17], the authors developed a framework for creating customized videos with both motion and structure of content retained. The approach enables users to customize video creation while keeping semantic consistency, which is an important step in personalized content generation.

The strengths of this framework are customizability with flexible options and precise motion retention, making it a useful tool for creators. It does need a lot of data to accommodate customized generation, which could be challenging for users who have limited resources.

In [18], the authors summarized the evolution and prospects of generative AI from engineering and societal angles. They highlight major landmarks and use with a general overview that is easily understandable for non-experts, who are looking for a basic familiarity with the subject. The review is informative but lacks a detailed exploration of technical implementations, so advanced readers might find themselves desiring detailed information in certain aspects.

In [19], the authors introduce a building-block-style text-to-video system that supports modular content creation. This novel system provides a new compositional form for generating video elements from text, enabling modular control of video content. The value of their work is that it is highly flexible, as users can readily control video components. Nonetheless, the system could struggle to ensure coherence between segments, which might degrade the coherence of generated videos as a whole.

The authors [20] investigate the use of generative AI in practical settings. With an emphasis on ethical design and user-centered considerations, the work highlights the wider ramifications of integrating generative AI across multiple domains. It offers insightful information about the responsible application of generative AI in real-world scenarios. Its limited technical depth with regard to model architectures and methodologies, however, may make it less helpful for researchers looking for in-depth implementation strategies.

In [21] ,the authors investigated the ways in which generative AI is revolutionizing creative industries, especially media and film. The research shows how tools are influencing traditional content creation and provides useful insights into changing workflows. The research is useful to industry practitioners, but it is not technically analyzed and is released in a non-tier journal, which can impact its validity within academe.
The authors [22] charted the evolution of generative AI based on a systematic review of published work. Their trend-categorization of generative models is a helpful perspective on the body of research that simplifies entry into the field for newcomers. The review is not so much about technical comparisons or innovations, however, which may restrict depth for senior researchers.

In [23], the authors integrated GANs with image segmentation methods for user-guided manipulation. Their approach enables targeted editing with segmented references to control generation, enhancing control and precision in image changes. The strength of this research is that it facilitates greater user interaction with generative models. Nevertheless, the success of their scheme depends significantly on the quality of segmentation, and the latter can be highly varied in different applications.

In [24], the authors explored the vulnerabilities of text-to-image models with GAN and CLIP to adversarial attacks. Their paper identifies flaws in model robustness, especially for slight input perturbations that trigger sizeable output changes. The research raises awareness about security threats to text-to-image systems, an important consideration for the development of more resistant models. Yet, the focus of their research is on one type of adversarial threat, and they may not be addressing the complete set of exposures inherent in generative models.

III. SUMMARY

The literature survey navigates prominent issues, developments, and advances in generative AI, with a special emphasis on text-to-video and text-to-image synthesis. Several studies [8, 16, 20]) attempt to improve quality and user controls on generation processes through LLM guidance, latent diffusion, and modular content structuring. Some others ([15, 23]) review model architectures and trends to facilitate newcomers and researchers in grasping the ideas behind diffusion and the evolution of generative models. Novel contributions such as personalized video synthesis [18], motion flow-guided video creation [17], and user-guided image editing via segmentation [24] indicate an interest in customizability and semantic consistency. On the other hand, [25] and [21] caution towards vulnerability and ethical issues of such models, putting emphasis on robustness and user experience. In all, while all these forms a solid base towards advancing the state of the art, there remain obvious limitations related to computational expense, data, generalization of motion, and real-world applicability.

1. TABLE
Taxonomy of literature review based on generative AI

Author, Year	Methodology	Contribution	Advantages	Limitations
M. Waseem et al., 2025	Combined LLM guidance with U-Net for T2V.	Improved video quality and relevance to text.	High fidelity and context awareness.	Requires heavy computation and data.
Zhang et al., 2024	Combined pixel and latent diffusion for T2V.	Hybrid model for improved video quality.	High fidelity, better consistency.	Complex and harder to optimize.
Kim et al., 2020	Stepwise generation: Text to Image then to Video via evolution generator	Early staged approach to T2V.	Simplifies generation process.	Outdated, lacks diffusion methods.
Ni et al., 2023	Used motion flow to create videos from images.	Produced realistic motion in video generation	Consistent and smooth output.	Struggles with complex motion.
Wu et al., 2025	Created a system for personalized video content.	Enabled user-specific video generation.	Flexible and detailed control.	Needs large training datasets.
Kilinç & Keçecioğlu, 2024	Reviewed GenAI's growth from broad perspectives.	Gave a general understanding of the field.	Easy for beginners to grasp.	Not technical in depth.
Ye Tian et al., 2024	Built a modular T2V system with block-wise control.	Allowed easy editing of video parts.	Supports modular design.	May affect overall video flow.
Yuan Sun et al., 2024	Analyzed GenAI in real-world use via UX view.	Highlighted user needs and ethics.	Focus on real-world design.	Light on technical detail.
Ketan Totlani, 2023	Studied GenAI's role in creative industries.	Showed how AI changes media workflows.	Industry-focused insights.	Over – reliance on automation.
García-Peñalvo & Vázquez-Ingelmo, 2023	Reviewed GenAI trends and categories.	Helped map research directions.	Easy overview for newcomers.	Lacks technical comparisons.
Watanabe et al., 2023	Merged GANs with segmentation for editing.	Enabled guided image changes.	Strong user control.	Depends on segmentation quality.
Chanakya et al., 2024	Explored adversarial risks in T2I models.	Flagged model weaknesses to small attacks.	Highlights security issues.	Narrow focus on one threat type.

IV. CONCLUSION

Modern generative AI models for creating images, videos, and other media have become increasingly accessible, thanks to open-source initiatives like Stability AI. The platform provides comprehensive tools for training, fine-tuning, and inference, alongside implementations of advanced architectures such as Stable Diffusion and other diffusion-based models. By making state-of-the-art generative AI technology, the platform empowers researchers and developers to push the boundaries of creativity and innovation across a wide range of applications. Emphasizing both high performance and scalability, the platform also advocates for ethical and responsible use of AI, addressing important societal concerns. Overall, an open-source effort plays a crucial role in advancing artificial intelligence by enabling the creation of high-quality synthetic content while fostering a collaborative and conscientious research environment

V. ACKNOWLEDGMENT

Would like to express our gratitude to the Department of Computer Science and Engineering, JNNCE for their support and encouragement.

VI. REFERENCES

[1] R. Togo, M. Kotera, T. Ogawa, and M. Haseyama, Text-Guided Style Transfer-Based Image Manipulation Using Multimodal Generative Models, IEEE Access, vol. 9, pp. 53870–53881, Mar. 2021

[2] Y. Peng, C. Zhao, H. Xie, T. Fukusato, and K. Miyata, Sketch-Guided Latent Diffusion Model for High-Fidelity Face Image Synthesis, IEEE Access, vol. 12, pp. 139–150, Dec. 2023

[3] W. Xiang, S. Xu, C. Lv, and S. Wang, A Customizable Face Generation Method Based on Stable Diffusion Model, IEEE Access, vol. 12, pp. 159000–159010, Dec. 2024

[4] S. K. Alhabeeb and A. A. Al-Shargabi, Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction, IEEE Access, vol. 12, pp. 29500–29515, Feb. 2024

[5] H. Kim, J.-H. Choi, and J.-Y. Choi, A Novel Scheme for Generating Context-Aware Images Using Generative Artificial Intelligence, IEEE Access, vol. 12, pp. 29800–29812, Mar. 2024

[6] J. Xing, M. Xia, Y. Liu, Y. Zhang, Y. Zhang, Y. He, H. Liu, H. Chen, X. Cun, X. Wang, Y. Shan, and T.-T. Wong, Make-Your-Video: Customized Video Generation Using Textual and Structural Guidance, IEEE Transactions on Visualization and Computer Graphics, vol. 31, no. 2, pp. 234–248, Feb. 2025

[7] V. Arkhipkin, Z. Shaheen, V. Vasilev, E. Dakhova, K. Sobolev, A. Kuznetsov, and D. Dimitrov, ImproveYourVideos: Architectural Improvements for Text-to-Video Generation Pipeline, IEEE Access, vol. 12, pp. 31560–31574, Jan. 2025

[8] M. Waseem, M. U. G. Khan, and S. K. Khurshid, LCGD: Enhancing Text-to-Video Generation via Contextual LLM Guidance and U-Net Denoising, IEEE Access, vol. 13, pp. 41234–41248, Mar. 2025

[9] Doyeon Kim, Donggyu Joo, Junmo Kim, "TiVGAN: Text to Image to Video Generation With Step-by-Step Evolutionary Generator," IEEE Access, vol. 8, pp. 150305–150315, Aug. 2020. doi: 10.1109/ACCESS.2020.3017881

[10] Haoxian Zhang, Ronggang Wang, Yang Zhao, "Multi-Frame Pyramid Refinement Network for Video Frame Interpolation," IEEE Access, vol. 7, pp. 129255–129265, Sept. 2019. doi: 10.1109/ACCESS.2019.2940510

[11] Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., & Shan, Y. (2024). VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024

[12] Jin, P., Li, H., Cheng, Z., Li, K., Ji, X., Liu, C., Yuan, L., & Chen, J. (2023). DiffusionRet: Generative Text-Video Retrieval with Diffusion Model. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 2023

[13] Zhang, D. J., Wu, J. Z., Liu, J. W., Zhao, R., Ran, L., Gu, Y., Gao, D., & Shou, M. Z. (2024). Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation. International Journal of Computer Vision (IJCV)

[14] P. Cao et al., "Controllable Generation with Text-to-Image Diffusion Models: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, early access, 2024, doi: 10.1109/TPAMI.2024.3392473

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695

[16] B. Ni, Y. Yang, C. Lu, W. Huang and T. Xiang, "Conditional Image-to-Video Generation with Latent Flow Diffusion Models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19720–19729

[17] Y. Wu et al., "CustomCrafter: Customized Video Generation with Preserving Motion and Concept Composition Abilities," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 4, 2025, pp. 4384–4392

[18] H. K. Kılınç and Ö. F. Keçecioğlu, "Generative Artificial Intelligence: A Historical and Future Perspective," Academic Platform Journal of Engineering and Smart Systems (APJESS), vol. 12, no. 2, pp. 47–58, 2024

[19] Y. Tian et al., "VideoTetris: Towards Compositional Text-to-Video Generation," in Advances in Neural Information Processing Systems (NeurIPS), vol. 37, 2024

[20] Y. Sun et al., "Generative AI in the Wild: Prospects, Challenges, and Strategies," in Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI), ACM, 2024

[21] K. Totlani, "The Evolution of Generative AI: Implications for the Media and Film Industry," International Journal for Research in Applied Science & Engineering Technology, vol. 11, no. 6, pp. 433–439, 2023

[22] F. J. García-Peñalvo and A. Vázquez-Ingelmo, "What Do We Mean by GenAI? A Systematic Mapping of the Evolution, Trends, and Techniques Involved in Generative AI," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 4, pp. 7–17, 2023

[23] Y. Watanabe, R. Togo, K. Maeda, T. Ogawa and M. Haseyama, "Text-Guided Image Manipulation via Generative Adversarial Network With Referring Image Segmentation-Based Guidance," IEEE Access, vol. 11, pp. 42534–42545, 2023, doi: 10.1109/ACCESS.2023.3269847