2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR) | 979-8-3503-5142-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/MIPR62202.2024.00018

10x Future of Filmmaking empowered by AIGC

Haohong Wang, Daniel Smith, and Malgorzata Kudelska
TCL Research America
San Jose, USA
{haohong.wang, daniel.smith, malgorzata.kudelska}@tcl.com

***Abstract*—** In this position paper, we present a vision for the future of filmmaking, driven by the emergence of generative AI technology. While the AI workflow for filmmaking remains in its infancy, recent advancements in diffusion models and the emergence of AI tools such as Midjourney, Runway, Pika, SORA, and LUMA are profoundly inspiring. The integration of AI into filmmaking endeavors, exemplified by projects like *Our T2 Remake* and *Next Stop Paris*, has yielded unprecedented impacts. This paper meticulously examines the challenges currently confronting AI models and proposes temporary solutions to surmount these obstacles in the filmmaking process. Furthermore, it demonstrates the workflow of the film "Next Stop Paris," illustrating how these integrated AI modules can collaborate efficiently to produce short films despite technical limitations in the early days. We foresee a future akin to Silicon Valley's technology incubation, where intellectual property (IP) incubation thrives in Hollywood. This initiative supports our vision of catalyzing 10x growth in the filmmaking industry.

***Keywords*—** *AI, generative AI, AIGC, diffusion model, filmmaking, Next Stop Paris.*

I. Introduction

With the latest advances in generative AI, also known as AIGC (Artificial Intelligence Generated Content), the timing is ideal for applying the 10x thinking or 10x growth strategy within the filmmaking industry. Embracing 10x thinking allows filmmakers to harness AI-driven tools and techniques to revolutionize their production processes, streamline workflows, and push the boundaries of creativity. This approach empowers them to envision projects on a scale previously unimaginable, explore new storytelling formats, and reach global audiences in groundbreaking ways. As AI technologies continue to advance rapidly, they offer unprecedented opportunities for innovation and efficiency in content creation.

It is essential to grasp the profound implications of the impending seismic shift in the filmmaking landscape propelled by AIGC technology. This transformative wave will indeed restructure conventional job roles while simultaneously birthing a multitude of new opportunities centered around AI-driven processes. Traditional facets of filmmaking, such as set construction, props, costume design, and labor-intensive tasks like editing, visual effects, and sound design, may experience a decline in demand. However, this evolutionary transition heralds a surge in the demand for specialized skills tailored to AI integration, digital asset creation, and automated editing mechanisms. For instance, professions like AI animators, tasked with managing AI tools and crafting immersive virtual landscapes, will become pivotal. Additionally, project managers will play a crucial role in streamlining production and post-production workflows. Moreover, as filmmaking capabilities become more accessible, we anticipate an era akin to the technology incubation witnessed in Silicon Valley, where intellectual property (IP) incubation will flourish in Hollywood. This trend promises writers, directors, and producers unprecedented access to AI-driven technologies, expediting the realization of their creative visions and IP projects within significantly shortened timeframes.

This convergence of AI and filmmaking not only brings efficiency gains but also opens new avenues for creative expression. By harnessing AI's prowess, filmmakers can delve into realms previously inaccessible, fostering innovation and propelling the industry into an era of unprecedented artistic exploration and technological advancement. Thus, while the transition may entail the restructuring of conventional job roles, it simultaneously ushers in a promising landscape brimming with opportunities for those adept at navigating the intersection of AI and cinematic creativity.

In recent years, numerous AI initiatives [1-10] have been explored in the realm of filmmaking automation. By 2023, AI tools such as Runway [11] and Pika [12] emerged, empowering users to create a few seconds of photorealistic professional quality video content with text or image prompts. In February 2024, OpenAI announced Sora [13], capable of generating longer videos (up to 60 seconds) with high quality. As of the paper's submission date, Sora remains unreleased and unavailable to the public. Nonetheless, these efforts instill strong confidence in this direction, and the tools establish a solid foundation for us to explore the future of filmmaking workflows using AIGC.

In this paper, we outline a vision of a 10x future for filmmaking to be achieved through the integration of AIGC technologies. This vision encompasses a comparison of the production pipelines between the proposed future AIGC workflow and traditional methodologies. Additionally, we address the challenges associated with employing current state-of-the-art AIGC technology to establish this workflow. Finally, we present a demonstration of the transitional workflow that we have utilized to produce one of the world's first AIGC love stories entitled Next Stop Paris.

2770-4319/24/$31.00 ©2024 IEEE
DOI 10.1109/MIPR62202.2024.00018

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO PARANA. Downloaded on November 05,2025 at 19:04:43 UTC from IEEE Xplore. Restrictions apply.

II. AIGC BREAKTHROUGHS RELEVANT TO FILMMAKING

A. Diffusion Models

Diffusion models represent a class of generative models that have gained prominence in recent years [20], [24], They experience rapid growth and are applied in various domains such as text-to-image generation, image-to-image generation, video generation, and 3D synthesis. The emergence of tools like DALL-E 2, Stable Diffusion [15], Midjourney [14], and Google's Imagen [34] has democratized machine learning, empowering users to create diverse images simply from text prompts. At their core, Diffusion models, learn to reverse a process of adding noise to training data, generating coherent images from noise. Through training, diffusion models learn to remove noise from images, using this denoising process to generate realistic images from random seeds. In other words, trained diffusion models can start with a random noise image and some conditioning information (e.g., a text description of the desired image) and can iteratively "denoise" the input signal, ending with a realistic output image.

Diffusion models have traditionally relied on U-Net architectures, which sequentially encode input images into lower-dimensional representations and subsequently decode them back to the original pixel space. While the original U-Net [29] only used ResNet blocks, most diffusion models interleave them with Vision Transformer blocks in each layer. Additionally, purely Vision Transformer-based diffusion models have emerged as alternatives to U-Net architectures, demonstrating distinct advantages such as adaptability in generating videos of varying lengths [23],[13],[27],[30]. Latest advancements have expanded diffusion models to incorporate video generation [25], offering the potential to revolutionize content creation. However, it introduces new challenges like ensuring spatial and temporal consistency, managing computational costs, and generating long video sequences.

To achieve temporal consistency, models need to share information across frames, often involving 3D architectures or factorized approaches to mitigate computational costs, while pre-processed features like depth estimates guide the denoising process for improved results. Typically, modifications are made to the self-attention layers within the U-Net architecture, including using temporal attention [10], full spatio-temporal attention [6], causal attention, and sparse causal attention [26]. Each form varies in computational demand and motion capture capability.

In the realm of filmmaking, video length poses a significant challenge. While short clips suffice for trailers or commercials, they fall short for full-length movies. A recent breakthrough, SORA from OpenAI [13],[30], represents the state-of-the-art in this field. It excels by producing videos up to a minute in length, all while preserving visual fidelity and staying true to the user's input. Another crucial challenge is achieving fine-grained control over content and motion synthesis. Human animation generation plays a pivotal role in maintaining consistent characters between scenes, thereby enhancing immersion and storytelling coherence. Techniques leveraging reference images and motion guidance specific to humans enable direct human animation video generation [9],[21],[22], facilitating seamless character continuity throughout a film. An extensive review on video diffusion models can be found in[28].

B. AIGC tools available to use

Among AI content generation tools, a distinction can be made between those with open-source models and those with models unavailable to the public. The primary advantage of the former group lies in the opportunity to develop proprietary tools and workflows built upon the model, enabling customization tailored to the specific requirements of filmmaking.

Popular AIGC tools include, but not limited to:

- *Midjourney* [14], which is an advanced AI program that generates images from natural language descriptions. Accessed through a Discord bot, users input prompts to receive sets of images, facilitating rapid prototyping.
- OpenAI *DALL·E 2* and its successor *DALL·E 3* [31]. DALL·E 3 is constructed on ChatGPT, enabling users to utilize ChatGPT for brainstorming and refining prompts.
- *Stable Diffusion* [15] - In August 2022, Stability AI released Stable Diffusion, an open-source diffusion model similar to DALL·E 2 and Google's Imagen [34]. Since then, they launched many new models including Stable Video Diffusion for video generation and Stable Video 3D that enables creating a 3D video of an object. By sharing the source code and model weights, Stability AI has made the model accessible to the wider AI community. Its availability of source code and model weights on GitHub enables users to utilize, fine-tune the models and develop various extensions and improvements. Furthermore, this open-source approach allows for customization tailored to specific needs, such as those required in filmmaking workflows. Numerous additional tools have been developed based on Stable Diffusion models, including GUI tools such as Stable Diffusion WebUI (AUTOMATIC1111 [33]) or ComfyUI [32]. They offer user-friendly interfaces for designing and executing Stable Diffusion pipelines. ComfyUI gives great flexibility by using a graph/nodes/flowchart-based approach. These tools have fostered a vibrant community that contributes by integrating various additional features like inpainting, super resolution, and many generation guidance techniques, enabling users to exert greater control over the generation of images or videos.
- *Runway* [11] is a tool that enables AI video generation. Their latent video diffusion model generates novel videos based on provided structural and content information. Structural consistency is maintained by conditioning on depth estimates, while content is governed by images or natural language prompts. The tool enables users to incorporate horizontal and vertical motion, camera roll, and zoom effects into their animations, thereby enriching the cinematic experience. It also provides a motion brush to add movement to specific areas of the animated scene.
- *Leonardo.ai* [16] offers image and video generation with text or image input, real-time canvas editing, 3D texture generation, one-click video asset creation, custom model

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO PARANA. Downloaded on November 05,2025 at 19:04:43 UTC from IEEE Xplore. Restrictions apply.

training, and the ability to use negative prompts to guide the generation process.

- *Pika [12]*, a free AI tool that can generate videos both based on text or image prompts
- *Sora [13]*, a new state-of-the-art diffusion model utilizing a transformer architecture, can generate videos up to a minute long while maintaining visual quality and adherence to the user's prompt. It is not yet publicly available, but OpenAI has given access to a group of professional artists and filmmakers to see what they could create.
- *Viggle [17]*, an AI engine that automates the creation of 3D character videos. It offers options for text-to-character, text-to-motion animation and generates character animations based on an input image of the character and a guiding motion video.

III. FUTURE PRODUCTION PIPELINE USING AIGC

The live-action filmmaking process follows three main phases: pre-production, production, and post-production. Pre-production involves the planning, scriptwriting, storyboarding, casting, location scouting, and set design. This phase ensures all elements are in place before shooting begins, ensuring a smooth production process. In the production phase of live-action filmmaking, the planned elements come to life. The director oversees the execution of the vision. Crew members handle lighting, sound, makeup, costuming, cinematography and a variety of other creative and physical production tasks. The film post-production phase involves the refining and assembling of raw footage into a polished final product. In editing, the visual segments are selected, arranged, and trimmed to construct the narrative. Visual effects are added to enhance scenes, and color correction is performed to ensure consistency and achieve the desired visual style. Sound design, including dialogue editing, sound effects, and music editing, creates the sound experience. Once all elements are integrated, the final film undergoes quality control checks before distribution.

3D Animated films follow a similar three-stage process. The pre-production process includes scriptwriting, concept development, character design, and storyboarding. The visual style, character personalities, and overall narrative structure are created before animation begins. Voice casting and recording also take place during this phase. In production, animators bring the planned elements to life. 3D artists craft digital models of characters, props, and environments based on pre-production designs. Textures are added to the models to achieve the desired visual effect. In rigging, a digital skeleton (or rig) is attached to the 3D models, enabling animators to pose and animate them realistically. In the animation stage, characters are brought to life through movement and performance. Lighting artists then set up virtual lights in each scene to illuminate characters and environments, creating the desired mood and atmosphere. Finally, the rendering process generates the final images or frames of the film from the 3D scene. The post production tasks in a 3D animated film are similar to a live-action film. These include compositing, editing, visual effects, sound design, and color grading/correction.

<table>
<tr><th>Production Pipelines</th><th colspan="2">Pre-Production</th><th>Production</th><th colspan="2">Post-Production</th></tr>
<tr><td>Animated film</td><td rowspan="3">Idea & Story Creation
Script Writing
Storyboarding</td><td>Voice Acting
Animatic</td><td>Modeling
Texturing
Rigging
Animation
Lighting
Rendering</td><td rowspan="3">Editing
VFX
Color correction
Sound design/Music/Mix</td><td>Compositing</td></tr>
<tr><td>Live action film</td><td>Location Scouting
Casting
Set Design & Construction
Costume design</td><td>Lighting/Rigging
Hair/Make-up/Costuming
Set Dressing
Props
Blocking
Cinematography
VFX/Sound</td><td></td></tr>
<tr><td>AIGC film</td><td>Voice Acting
Character Design with AI
Scene Design with AI</td><td>Animation with AI
Compositing with AI</td><td></td></tr>
</table>

Figure. 1. The comparison of production pipelines of three types of titles: (1) Animated film, (2) Live action film; (3) AIGC film

In a future in which films are produced using AIGC technology, the production workflow would undergo a radical transformation compared to traditional filmmaking processes. As illustrated in Fig. 1, while the pre-production and post-production phases of AIGC films closely resemble those of traditional filmmaking, the production phase of AIGC is streamlined and the workflow is outlined as follows:

Character Design with AI: AI systems are employed in the design and generation of virtual characters, creatures, and entities, covering aspects such as their appearance, voice, anatomy, and behavioral traits. Users can influence the design outcome by providing prompts or guides in either text, image or other format.

Scene Design with AI: AIGC algorithms are utilized in the design and generation of virtual environments, landscapes, and architectural structures, based on the script and narrative requirements. Users can control the outcome of the design simply by providing prompts or guides in either text, image or other formats.

Animation with AI: Starting from a text,still image, or video prompt, AI animates characters and objects, imbuing them with lifelike movements, expressions, and interactions. Users can control the outcome of the animation simply by providing prompts or guides in either text, image or other format.

Composition with AI: AIGC algorithms autonomously compose shots, camera movements, and visual compositions based on cinematic principles, storytelling goals, and emotional cues. Leveraging reinforcement learning and predictive modeling, AI dynamically adjusts camera angles, framing, and lighting to enhance narrative impact and audience engagement.

In this speculative future scenario, the production workflow of a film using AIGC technology would be characterized by unprecedented levels of automation, creativity, and innovation. While human input will still play a role in setting objectives, defining parameters, and guiding the overall creative direction, most production tasks would be streamlined, enhanced, or automated by AI systems. This may lead to the emergence of entirely new forms of storytelling and visual expression. Addressing the ethical considerations regarding AI-generated content, creative authorship, and cultural representation would be paramount, requiring careful scrutiny and responsible stewardship by filmmakers, producers, and other stakeholders.

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO PARANA. Downloaded on November 05,2025 at 19:04:43 UTC from IEEE Xplore. Restrictions apply.

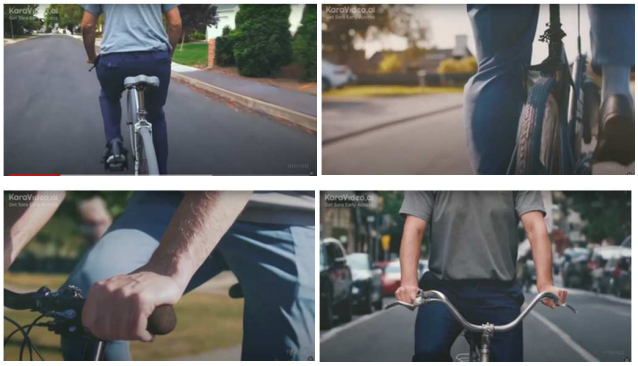

Figure 2. Selected frames shown from *air head, Made by shy kids with Sora,* with inconsistency issues

IV. CHALLENGES FACED BY STATE-OF-THE-ART AIGC

As mentioned in section 2.B, in the realm of AIGC, a multitude of tools and models have surfaced, showcasing promising capabilities for content creation. Yet, amidst this proliferation, several challenges must be confronted to harness Generative AI's full potential in filmmaking.

A. Consistency of generated characters and backgrounds

In filmmaking, it's crucial for characters to maintain their identity consistently across different scenes, even when there are changes in clothing or hairstyle. Additionally, generating images of the same background from various angles and positioning figures in photorealistic poses are essential requirements. Despite the capabilities of several renowned tools, such as Midjourney or Leonardo.ai, in generating impressively realistic images, they face challenges in ensuring the consistent appearance of characters across multiple scenes, even when provided with a sample input image as a reference. As depicted in Fig. 2, inconsistency issues are evident in the videos generated using Sora. Within the first 20 seconds of the video, noticeable changes occur in the character's shirt color, bicycle head style, and shoe color.

Within the realm of open-source tools harnessing Stable Diffusion models, various solutions have emerged to tackle this issue. Some strategies involve fine-tuning the Stable Diffusion model through techniques like Dreambooth [36], Low Rank Adaptation (LoRA) [37], or Textual inversion [38]. Others, such as IP-Adapter [39] and PhotoMaker [40], employ specialized adapters to enable image prompt functionality for pre-trained text-to-image diffusion models. In fine-tuned models, challenges arise in achieving a faithful generation of the desired character without compromising the quality of other elements, such as background colors and details. Conversely, when utilizing image prompting, the final image's resemblance to the input image often exhibits less accuracy.

Generating a specific character in a particular pose within a designated background becomes even more complex, especially when both the character and background are generated by separate fine-tuned models. In such scenarios, intricate workflows are necessary, incorporating setups like ControlNet [42] to direct the generation process with additional inputs such as extracted skeletons from OpenPose [41] for the character or Canny edges, depth maps, etc., for the background. Regarding background generation, challenges persist in achieving consistent geometry and detail, as well as in creating 3D models to guide image generation from different viewpoints.

Figure 3. The generated frames before and after camera movements

Lastly, achieving consistency and stability in generated animations involves addressing issues such as background flickering and character degradation over time. This includes animating characters within specified backgrounds and preserving frame quality in longer videos. Despite ongoing research, achieving seamless interaction between characters and backgrounds in AI-generated content remains a significant hurdle. In the demos provided by Sora [13], it asserts its capability to address the above challenges by effectively modeling short- and long-range dependencies. It maintains the appearance of characters and objects, even when occluded or transitioning off-screen, and can generate multiple shots of the same character within a single sample with consistent appearance throughout the video.

B. Controllability of generated camera movements

In the domain of AI-driven filmmaking, achieving versatility in camera movements (such as zooming and tilting) while maintaining coherence between scenes poses a critical challenge. Current AI-driven content generation tools encounter limitations in accurately replicating dynamic camera maneuvers, thus imposing significant constraints on filmmakers. Consistency between scenes is equally vital. To ensure a seamless flow, static images must be animated to maintain consistency in characters and backgrounds. However, existing solutions encounter challenges in this regard. For instance, AnimateDiff [18] utilizes specialized models called motion LoRAs to create preset camera movements. Nonetheless, this often results in significant changes in image details, affecting both characters and backgrounds (as shown in Fig. 3). Conversely, MotionCtrl [19] offers control over camera movements but lacks the same level of control over characters and other scene elements.

To surmount these challenges, we require solutions that offer simultaneous control over both camera movements and scene elements during animation. By refining motion models to synchronize camera movements with character actions, filmmakers can enhance their storytelling capabilities with greater precision. In the demonstrations provided by Sora, it showcases its ability to generate videos with dynamic camera motion. As the camera undergoes shifts and rotations, individuals and scene elements seamlessly traverse a three-dimensional space, ensuring a cohesive and immersive viewing experience.

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO PARANA. Downloaded on November 05,2025 at 19:04:43 UTC from IEEE Xplore. Restrictions apply.

C. *Generating animations with entity/character interaction or in high motion*

Generating animations involving entity or character interaction, or high-motion sequences, presents a significant challenge in the field of generative AI. This difficulty stems from the complexities involved in simulating realistic interactions, whether between a character and the background, two characters themselves, or objects in high motion (as depicted in Fig. 4, where the lady's hands are not depicted accurately or naturally). Currently, AI models often struggle to accurately capture the nuances of these interactions, resulting in animations that may appear unnatural or disjointed. Addressing this challenge is crucial for advancing the capabilities of generative AI in producing more immersive and believable animated content. To address these challenges, efforts are underway to integrate 3D models into generative AI systems, aiming to enhance the accuracy of object and human motion generation [43][44][22]. Additionally, developing solutions that address complex interactions with the environment and understand specific cause-and-effect relationships is paramount. Despite its apparent advancements in generating high-quality and realistic animations, even Sora [13] is reported to face limitations. While it excels in rendering convincing entity interactions, such as playful puppies, the current model exhibits weaknesses. Challenges arise in accurately simulating complex scene physics and comprehending cause-and-effect relationships. Additionally, spatial details and precise descriptions of events over time may pose significant obstacles.

V. NEXT STOP PARIS PRODUCTION

On April 12, 2024, TCL unveiled the trailer for its inaugural AIGC film, "Next Stop Paris." This announcement garnered significant attention from the press, sparking excitement among some individuals regarding the progression of AIGC technology into the filmmaking realm. While some enthusiasts are willing to embrace the inherent immaturity of AI technology, others remain skeptical about the outcomes of AIGC, citing the challenges highlighted in the previous *section*.

Figure 4. The generated frames in high motion

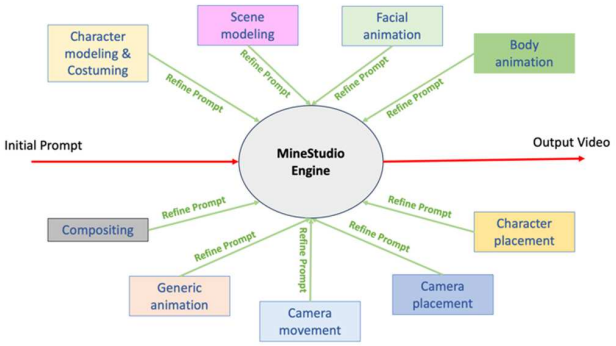

Figure 5. The conceptual I/O interface of MineStudio Engine used in film *Next Stop Paris* production

As depicted in Fig. 5, the production of "Next Stop Paris" was facilitated by an in-house AI solution named MineStudio. This platform enables users to input an initial text prompt to generate a video. Subsequently, users can utilize multiple refine prompts, such as text, images, poses, camera settings, motion, and other form factors. These refine prompts allow for adjustments to various aspects of the video, including characters and associated costumes, background scenes, facial and body animations, camera settings, and more, ultimately refining the generated video output. As indicated in [35], shots below 5 seconds are the most common elements in Animated films. In our AIGC filmmaking process, shots under 10-seconds comprise the core elements of the film. Special techniques are required to make sure the characters and background can cross multiple shots without inconsistency issues. Therefore, the units processed in MineStudio are quick shots spanning a few seconds. In this paper, the black box of MineStudio is opened (as depicted in Fig. 6) to illustrate how this system is constructed to adapt effectively to the rapid advancements in AI technology.

The system can be conceptualized as comprising three interconnected subsystems, facilitated by a prompt and guide connector. Firstly, an image diffusion model enables image generation with text prompts, pose, clothing style, and camera angle guidance. Some artists may prefer to utilize tools such as Pika or Runway to generate videos from well-designed, high-quality images. Secondly, a character animation flow is implemented to manage facial animation, body animation, and the refinement process for both facial and body elements. This flow can accept inputs such as live-action model videos or character models, along with text or image prompts. Lastly, a background video processing flow is established to handle background settings, camera movements, video composition, and quality refinement, along with frame interpolation. This flow takes inputs such as background scene models, camera angle guides, and text or image prompts to complete the processing.

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO PARANA. Downloaded on November 05,2025 at 19:04:43 UTC from IEEE Xplore. Restrictions apply.

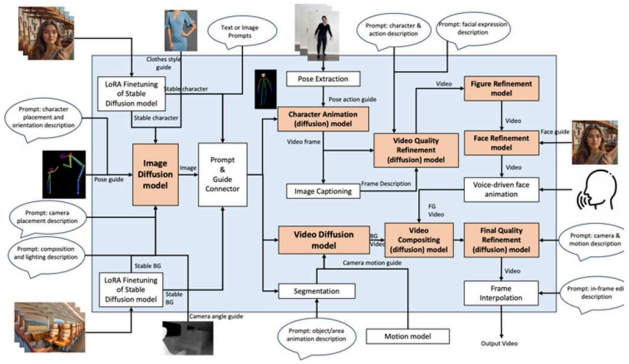

Figure 6. The MineStudio workflow used in film *Next Stop Paris* production

The open interface design of the three key modules, image diffusion model, character animation (diffusion) model, and video diffusion model, allows the system to integrate multiple available AI tools, such as Midjourney, Stable Diffusion, Pika, Runway, Viggle, and potentially SORA in the future. This integration complements TCL's in-house AI models, maximizing performance within the platform. The prompt and guide connector facilitates the transmission of various types of text or image prompts, as well as diverse guides for face, pose, clothing style, background, camera, and more, to the video generation modules. This empowers animators with significant control over the tailoring of the videos they wish to produce. Consequently, users' demands or prompts can influence the effect of the generated video by being passed to the relevant modules. Additionally, with LoRA models and quality refinement modules, inconsistency issues can be mitigated when transitioning across modules or shots. For more AIGC video clips, readers can visit the https://tcltv.plus/ website.

VI. CONCLUSION

This article serves as a position paper, illustrating the impact of ongoing AIGC breakthroughs on the film industry through the introduction of a revolutionary filmmaking production pipeline. It examines the challenges posed by AIGC technology, particularly in meeting the high-quality standards expected in television shows and films. By revealing the workflow behind one of the first AIGC films, "Next Stop Paris," the paper compellingly demonstrates how multiple AI tools can be seamlessly integrated into an efficient pipeline. This integration empowers artists to enhance their productivity and capabilities, thereby accelerating the production process. As filmmakers increasingly adopt AI technology, the paper suggests that the future of filmmaking, characterized by a 10x increase in efficiency, can be realized in an unprecedentedly short period.

ACKNOWLEDGMENT

We would like to extend our gratitude to all the team members who contributed to the AIGC film, "Next Stop Paris," including Chris Regina, Catherine Zhang, Allan He, Julie Rothman, and many artists, researchers, and engineers behind the scenes.

REFERENCES

[1] Z. Yu, H. Wang, A. K. Katsaggelos, J. Ren, "A Novel Automatic Content Generation and Optimization Framework," *IEEE Internet of Things Journal*, 10(14): 2338-12351, 2023.

[2] Z. Yu, X. Wu, H. Wang, A. K. Katsaggelos, J. Ren, "Automated Adaptive Cinematography For User Interaction in Open World," *IEEE Transactions on Multimedia*, to appear.

[3] X. Wu, H. Wang, A. K. Katsaggelos, "The secret of immersion: actor driven camera movement generation for auto-cinematography," *arXiv preprint arXiv:* 2303.17041, 2023.

[4] K. Jorgensen, H. Wang, M. Wang, "From Screenplay to Screen: A Natural Language Processing Approach to Animated Film Making", in *Proc. IEEE ICNC 2023*, pp. 484-490, 2023.

[5] Z. Yu, C. Yu, H. Wang, J. Ren, "Enabling Automatic Cinematography with Reinforcement Learning," in *Proc. IEEE MIPR 2022*, pp. 103-108, 2022.

[6] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, et al, "Lumiere: A Space-Time Diffusion Model for Video Generation", *arXiv preprint arXiv*: 2401.12945, 2024.

[7] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, A. Germanidis, "Structure and Content-Guided Video Synthesis with Diffusion Models", in *Proc. ICCV 2023*, pp. 7346-7356, 2023.

[8] W. Peebles, S. Xie, "Scalable Diffusion Models with Transformers", in *Proc. ICCV 2023*, pp. 4195-4205, 2023.

[9] L. Hu, X. Gao, P. Zhang, K. Sun, B. Zhang, L. Bo, "Animate Anyone: Consistent and Controllable Image-to-Video synthesis for Character Animation", arXiv preprint arXiv: 2311.17117, 2023.

[10] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al, "Make-a-video: Text-to-video generation without text-video data". *arXiv preprint arXiv:*2209.14792, 2022.

[11] Runway, https://runwayml.com

[12] Pika, https://pika.art/home

[13] Sora, https://openai.com/sora

[14] Midjourney, https://www.midjourney.com/

[15] Stable Diffusion, https://stability.ai/

[16] Leonardo.ai, https://leonardo.ai/

[17] Viggle, https://viggle.ai/

[18] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, B. Dai, "AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning," *arXiv preprint arXiv:* 2307.04725, 2023.

[19] Z. Wang, Z. Yuan, X. Wang, T. Chen, M. Xia, P. Luo, Y. Shan, "MotionCtrl: A Unified and Flexible Motion Controller for Video," *arXiv preprint arXiv:* 2312.03641, 2023.

[20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022 pp. 10674-10685, 2022.

[21] Z. Xu, J. Zhang, J.H. Liew, H. Yan, J.W. Liu, C. Zhang, J. Feng, M.Z. Shou, "MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model," *arXiv preprint arXiv:* 2311.16498, 2023.

[22] S. Zhu, J. L. Chen, Z. Dai, Y. Xu, X. Cao, Y. Yao, H. Zhu, S. Zhu, "Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance," *arXiv preprint arXiv:* 2403.14781, 2024.

[23] W. Peebles and S. Xie, "Scalable diffusion models with transformers," *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

[24] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, 2020.

[25] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, D. J. Fleet, "Video Diffusion Models," In *Advances in Neural Information Processing Systems*, 2022.

[26] J. Z. Wu, Y. Ge, X. Wang, W. Lei, Y. Gu, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," *arXiv preprint arXiv:* 2212.11565, 2022.

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO PARANA. Downloaded on November 05,2025 at 19:04:43 UTC from IEEE Xplore. Restrictions apply.

[27] F. Bao et al., "All are Worth Words: A ViT Backbone for Diffusion Models," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, pp. 22669-22679, 2023.

[28] A. Melnik, M. Ljubljanac, C. Lu, Q. Yan, W. Ren, H. Ritter, "Video Diffusion Models: A Survey," *arXiv preprint arXiv:* 2405.03150, 2024.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference,* Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241, Springer, 2015.

[30] Y. Liu, K. Zhang, Y. Li, Z. Yan, C. Gao, R. Chen, Z. Yuan, Y. Huang, H. Sun, J. Gao, L. He and L. Sun, "Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models," *arXiv preprint arXiv:* 2402.17177, 2024.

[31] DALLE 3, https://openai.com/index/dall-e-3/

[32] ComfyUI, https://github.com/comfyanonymous/ComfyUI

[33] AUTOMATIC1111, https://github.com/AUTOMATIC1111/stable-diffusion-webui

[34] Imagen, https://imagen.research.google/

[35] Hah EJ, Schmutz P, Tuch AN, Agotai D, Wiedmer M, Opwis K. Cinematographic techniques in architectural animations and their effects on viewers' judgment. *International Journal of Design*. 2008 Dec 31;2(3).

[36] N. Ruiz, Y. Li, V. Jampani,Y. Pritch, M. Rubinstein and K. Aberman, "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," *arXiv preprint arXiv: 2208.12242*, 2023.

[37] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv: 2106.09685*, 2021.

[38] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik and D. Cohen-Or, "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion," *arXiv preprint arXiv: 2208.01618*, 2022.

[39] H. Ye, J. Zhang, S. Liu, X. Han and W. Yang, "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models," *arXiv preprint arXiv: 2308.06721*, 2023.

[40] Z. Li, M. Cao, X. Wang, Z. Qi, M. Cheng and Y. Shan, "PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding," *arXiv preprint arXiv: 2312.04461*, 2023.

[41] Z. Cao, G. Hidalgo, T. Simon, S. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *arXiv preprint arXiv: 1812.08008*, 2019.

[42] L. Zhang, A. Rao and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," *arXiv preprint arXiv: 2302.05543*, 2023.

[43] V. Voleti, C. Yao, M. Boss, A. Letts, D. Pankratz, D. Tochilkin, C. Laforte, R. Rombach and V. Jampani, "SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion," *arXiv preprint arXiv: 2403.12008*, 2024.

[44] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa and J. Malik, "Humans in 4D: Reconstructing and Tracking Humans with Transformers,", *arXiv preprint arXiv: 2305.20091*, 2023.

Authorized licensed use limited to: UNIVERSIDADE FEDERAL DO PARANA. Downloaded on November 05,2025 at 19:04:43 UTC from IEEE Xplore. Restrictions apply.