

Generative AI for Cel-Animation: A Survey

Yunlong Tang¹, Junjia Guo¹, Pinxin Liu¹, Zhiyuan Wang², Hang Hua¹, Jia-Xing Zhong³, Yunzhong Xiao⁴, Chao Huang¹, Luchuan Song¹, Susan Liang¹, Yizhi Song⁵, Liu He⁵, Jing Bi¹, Mingqian Feng¹, Xinyang Li¹, Zeliang Zhang¹, Chenliang Xu¹

¹University of Rochester, ²UCSB, ³University of Oxford, ⁴CMU, ⁵Purdue University

{yunlong.tang, jing.bi, mingqian.feng, chenliang.xu}@rochester.edu, zwang796@ucsb.edu, {jguo40, pliu23, lsong11, sliang22, xli190, zzh136}@ur.rochester.edu, jiaxing.zhong@cs.ox.ac.uk, {hhua2, chuang65}@cs.rochester.edu, yunzhonx@andrew.cmu.edu, {song630, he425}@purdue.edu

<https://github.com/yunlong10/Awesome-AI4Animation>

Abstract—Traditional Celluloid (Cel) Animation production pipeline encompasses multiple essential steps, including storyboard, layout design, keyframe animation, inbetweening, and colorization, which demand substantial manual effort, technical expertise, and significant time investment. These challenges have historically impeded the efficiency and scalability of Cel-Animation production. The rise of generative artificial intelligence (GenAI), encompassing large language models, multimodal models, and diffusion models, offers innovative solutions by automating tasks such as inbetween frame generation, colorization, and storyboard creation. This survey explores how GenAI integration is revolutionizing traditional animation workflows by lowering technical barriers, broadening accessibility for a wider range of creators through tools like AniDoc, ToonCrafter, and AniSora, and enabling artists to focus more on creative expression and artistic innovation. Despite its potential, issues such as maintaining visual consistency, ensuring stylistic coherence, and addressing ethical considerations continue to pose challenges. Furthermore, this paper discusses future directions and explores potential advancements in AI-assisted animation. For further exploration and resources, please visit our GitHub repository: <https://github.com/yunlong10/Awesome-AI4Animation>

Index Terms—Generative AI, Cel-Animation.

I. INTRODUCTION

Animation, as a powerful medium for storytelling and artistic expression, has evolved significantly over the past century. Celluloid (Cel) animation, established as a cornerstone of traditional animation in the early 1920s, has shaped the foundation of the modern animation industry through its distinctive frame-by-frame approach that combines artistic vision with technical precision. This methodology laid the foundation for modern animation while highlighting a fundamental tension: the challenges of maintaining high artistic quality while improving production efficiency: 1) *Time-intensive manual labor*: The creation of keyframes, inbetweening, and colorization often requires extensive manual effort. 2) *Technical complexity*: Coordinating multiple stages, from scripts to storyboard, from keyframe animation to coloring, involves a high degree of expertise and meticulous planning. 3) *Creativity constraints*: Repetitive and labor-intensive tasks can detract from the time and energy animators devote to higher-level creative decisions. These challenges pose barriers to efficiency and scalability, particularly in modern animation workflows where production timelines are tightening.

Recent advances in generative artificial intelligence (GenAI) have introduced transformative solutions to these challenges.

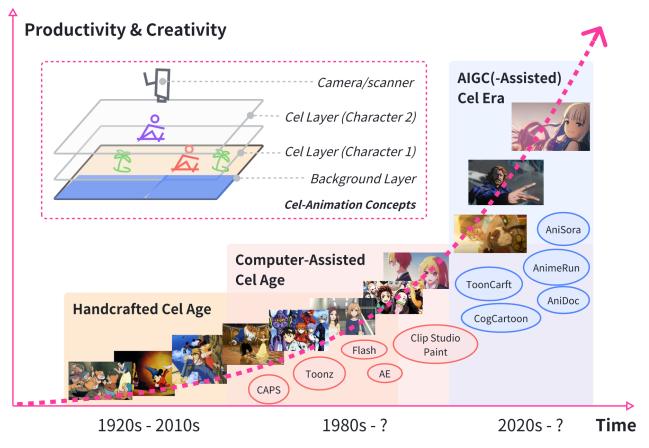


Fig. 1. The three major phases in Cel-Animation history: the Handcrafted Cel Age (1920s–2010s), the Computer-Assisted Cel Age (1980s–present), and the emerging AIGC Cel Era (2020s onward). A layered structure of Cel-Animation is also shown. The names of the works featured in the teaser can be found in Appendix A.

GenAI, including large language models (LLMs) [1]–[3], multimodal large language models (MLLMs) [4]–[9], and diffusion models [10]–[14], has demonstrated its ability to produce images, videos, and even animations autonomously or semi-autonomously. By automating repetitive tasks, such as generating inbetween frames or applying color schemes, GenAI enables animators to focus more on creative storytelling and artistic innovation. Its applications in various creative fields are expanding rapidly, showing promise in addressing the inherent challenges of Cel-Animation production.

To better understand how GenAI is transforming Cel-Animation, we examine its historical evolution through three major stages, as shown in Figure 1:

1) *The Handcrafted Cel Age (1920s-2010s)*: Cel-Animation emerged in the early 20th century as a groundbreaking art form that revolutionized animation production. The innovation of using celluloid sheets enabled animators to separate dynamic characters from static backgrounds, allowing for unprecedented depth in scenes and efficient reuse of background elements [15]. This process established fundamental techniques still used today: storyboard for narrative planning, keyframe and inbetween frame drawing for motion, and hand-painting cels for final

visuals. While masterpieces like *Snow White and the Seven Dwarfs* (1937) showcased the artistic potential, they also highlighted its limitations—the film required over 200,000 hand-drawn frames and three years to complete. The labor-intensive process of creating 24 frames per second demanded large teams of specialized artists, presenting significant challenges in maintaining consistency and scaling production.

2) *The Computer-Assisted Cel Age (1980s-present)*: The 1980s marked a significant shift in Cel-Animation with the introduction of digital tools that revolutionized traditional workflows. Disney's Computer Animation Production System (CAPS), developed in collaboration with Pixar, pioneered the automation of coloring and compositing while preserving the aesthetic of hand-drawn animation [16]. Films like *Beauty and the Beast* (1991) and *The Lion King* (1994) demonstrated CAPS' ability to enhance precision and complexity. Concurrently, Japanese studios adopted tools like Toonz [17], OpenToonz [18], Adobe Flash [19], and Clip Studio Paint [20] to optimize inbetweening and compositing while maintaining their signature hand-drawn style, as seen in *Nausicaä of the Valley of the Wind* (1984) and *Neon Genesis Evangelion* (1995). This era showcased contrasting strategies: U.S. studios continued to use full animation, while Japanese studios sufficiently utilized “limited animation” [21] further reduced frame counts and maximized resource efficiency, which brought significant productivity. While these digital tools enhanced efficiency and expanded creative possibilities, they primarily served as assistive technologies, with core artistic execution still heavily dependent on manual input.

3) *The AIGC Cel Era (2020s onward)*: The emergence of Artificial Intelligence Generated Content (AIGC) has redefined Cel-Animation. Unlike previous digital tools that mainly augmented manual processes, AIGC actively participates in technical execution, such as inbetween frame generation, complex effects creation, and stylistic consistency maintenance [22]–[24]. These tools reduce repetitive tasks, allowing animators to focus on creative decisions [25]. For instance, AniDoc employs video diffusion models to automate inbetweening and colorization in 2D animation [23], while ToonCrafter [24] efficiently handles exaggerated non-linear motions and occlusions in cartoons. Platforms like Storyboarder.ai streamline pre-production by generating storyboards using AI [26]. Experiments by Netflix Japan (*The Dog and the Boy*, 2023) and tools like CogCartoon [27] demonstrate AIGC’s potential to democratize animation by enabling practical visualization with minimal resources. Research in animation video generation models, such as AniSora [28], and innovations in generative storytelling [25], further highlight the potential for AIGC to revolutionize animation as an art form and a production methodology.

While AIGC has dramatically improved animation production efficiency and creative possibilities, significant challenges persist in maintaining visual consistency, and preserving stylistic coherence [22]–[24, 29]. As the field continues to evolve rapidly, understanding these challenges and opportunities requires a systematic examination of both traditional animation techniques and emerging AI technologies.

Recently, several surveys have explored related fields in

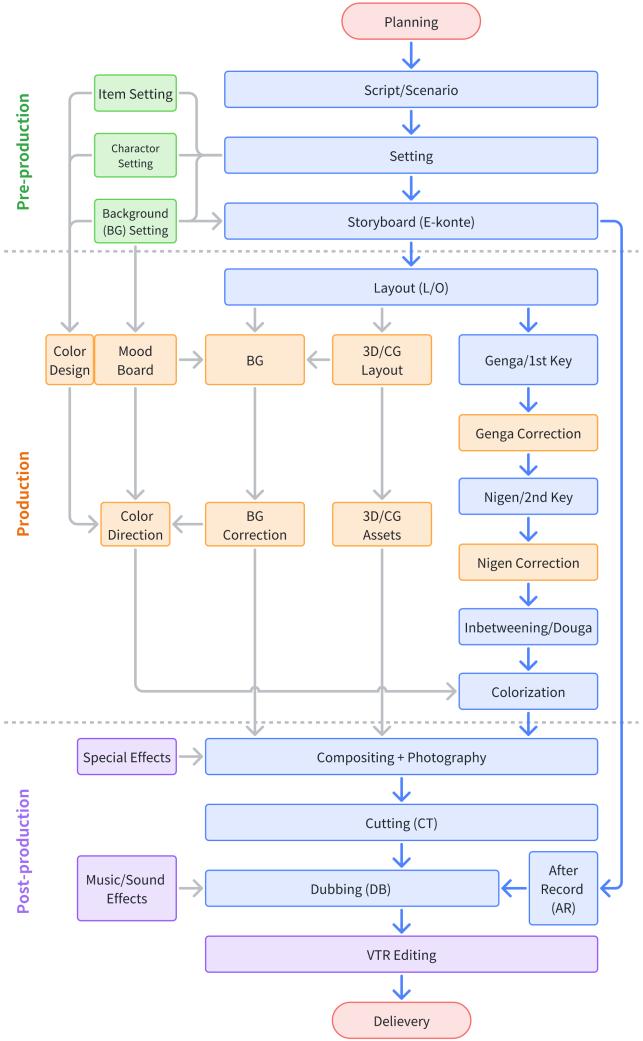


Fig. 2. Comprehensive workflow diagram of a traditional animation production pipeline, illustrating the stages of planning, pre-production, production, and post-production, with detailed processes such as scripting, setting, storyboarding, layout, keyframe animation, inbetweening, colorization, photography, dubbing, etc.

generative content creation. Li et al. [30] focus on long video generation, addressing challenges and methodologies for extended video content. Lei et al. [31] review human video generation, emphasizing realistic motion synthesis. Xing et al. [32] survey video diffusion models, highlighting their application to temporal coherence. Zhou et al. [33] explore generative AI and large language models in video generation, and Cho et al. [34] provide a comprehensive review of text-to-video generation. Zhao et al. [35] present a comprehensive survey on cartoon image processing, including a limited coverage of topics related to cartoon animation. In contrast, our work specifically targets Cel-Animation, systematically reviewing the integration of GenAI tools into traditional animation pipelines, analyzing its role across pre-production, production, post-production stages, and discussing future research directions, with particular attention to how these technologies are reshaping the balance between creativity and productivity in animation.

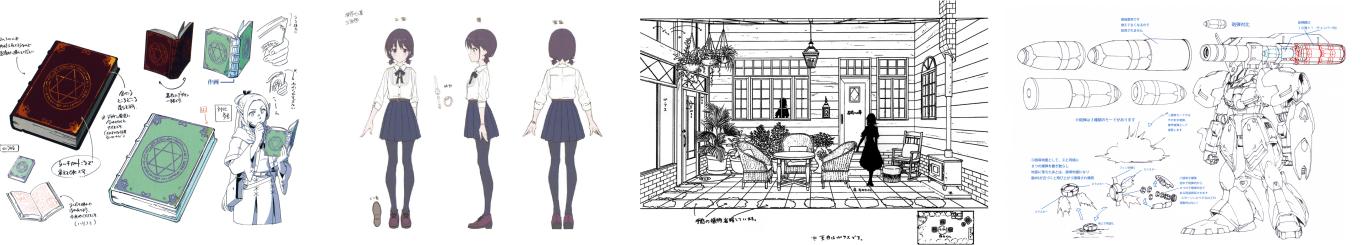


Fig. 3. Examples of settings (setteis) of items, characters, scenes, and mecha. The examples are from *Delicious in Dungeon* (2024), *Girls Band Cry* (2024), *Violet Evergarden* (2018), and *Gundam Build Divers* (2018), respectively.

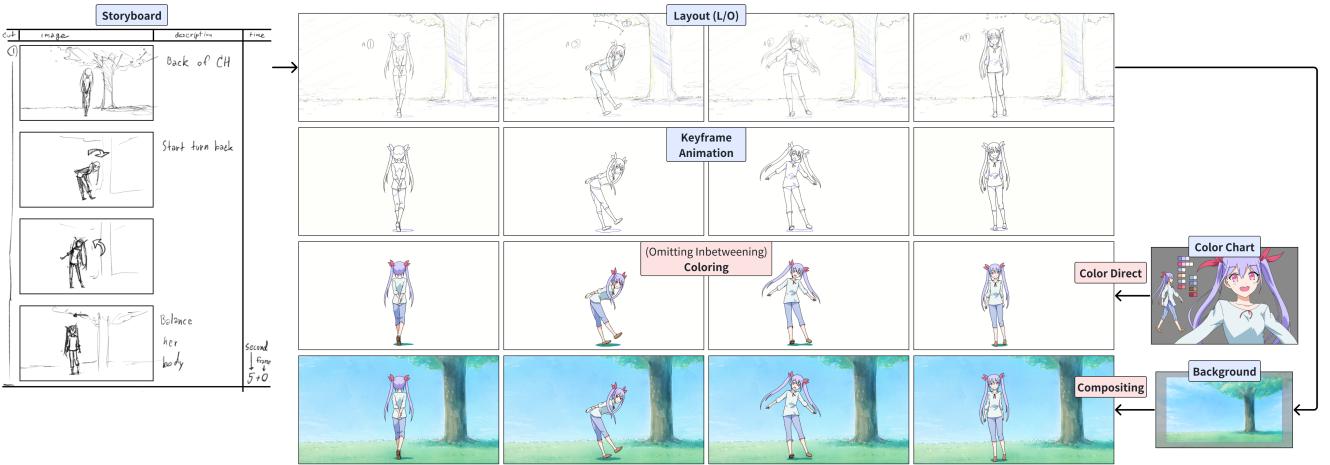


Fig. 4. A production example showing the transformation of a scene from storyboard to final compositing, demonstrating key stages including layout (L/O), keyframe animation, coloring, and background integration.

The remainder of this survey is structured as follows: in the Preliminary Section II, we provide an overview of the conventional Cel-Animation pipeline, breaking it down into pre-production, production, and post-production stages. In this section, we also offer foundational knowledge on GenAI, covering topics such as LLMs, LMMs, and diffusion models. Section III introduces the specific applications of GenAI in Cel-Animation, highlighting its role in enhancing efficiency and creativity. Based on the current research, Section IV discusses the limitations and future research directions, and Section V concludes the survey paper.

II. PRELIMINARY

A. Conventional Cel-Animation Pipeline

The Cel-Animation production pipeline is a meticulously structured process, encompassing multiple stages from conceptualization to final delivery. Each phase plays a vital role in shaping the quality and visual appeal of the final animation. This section provides a comprehensive overview of the conventional Cel-Animation workflow, which includes pre-production, production, post-production, and other auxiliary processes, as illustrated in Figure 2. More interpretation of terms can be found in Appendix C.

1) **Pre-production:** Pre-production sets the foundation for an animation project by defining its creative direction and technical framework. It is a highly collaborative phase that determines the overarching vision of the work.

- **Scripting:** The script serves as the blueprint for the story, defining the time, place, characters, events, and dialogue. It also specifies which props and items need to appear in the scenes, along with their respective colors. In cases of adaptations or original works, initial meetings involve directors, scriptwriters, and producers discussing the overall story structure, world-building, and required narrative elements. Decisions made at this stage influence the project's trajectory.

- **Setting:** Building upon the script, the setting defines the world in which the story takes place, encompassing both the physical and thematic environment. This phase involves conceptualizing characters, items, and background elements that visually represent the story's mood and themes (see Figure 3). ¹ They are used as templates by animators in order to stay on the model when drawing.

- **Storyboarding:** The storyboard translates the script into a visual sequence, determining shot composition, character acting, and scene timing. Directors and storyboard artists work closely during storyboard meetings to finalize the visual narrative. This step establishes the creative groundwork for subsequent production phases. An example of storyboard can be seen in Figure 4 on the left.

2) **Production:** Production is where pre-production ideas take physical form, transitioning from conceptualization to the creation of animation assets. Multiple teams collaborate

¹<https://setteidreams.net/settei/>

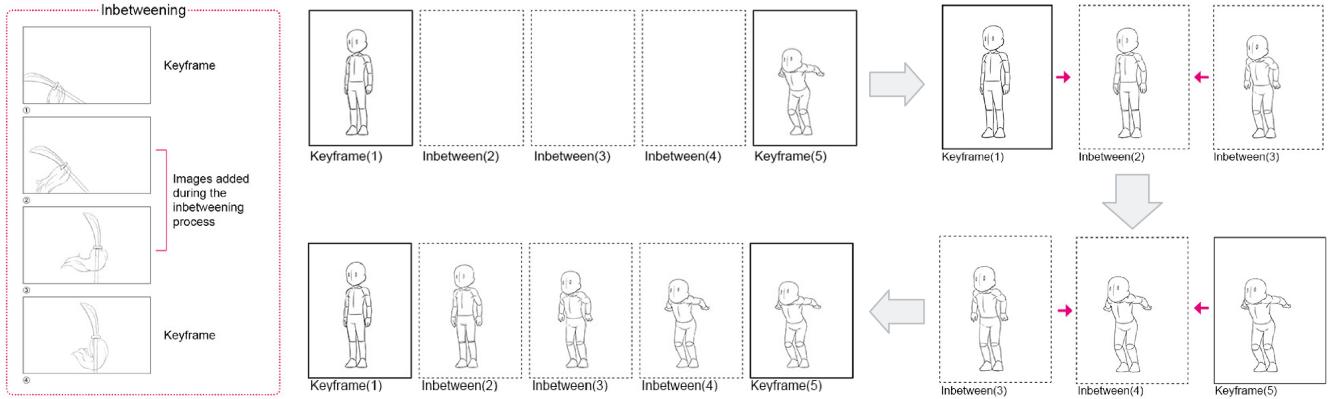


Fig. 5. The conventional process of inbetweening involves adding inbetween frames between every two keyframes to create smoother motion and transitions. In this example, Inbetween(3) is created first based on Keyframe(1) and Keyframe(5). Then, Inbetween(2) can be created based on Keyframe(1) and Inbetween(3). Similarly, Inbetween(4) can be obtained in the same way.

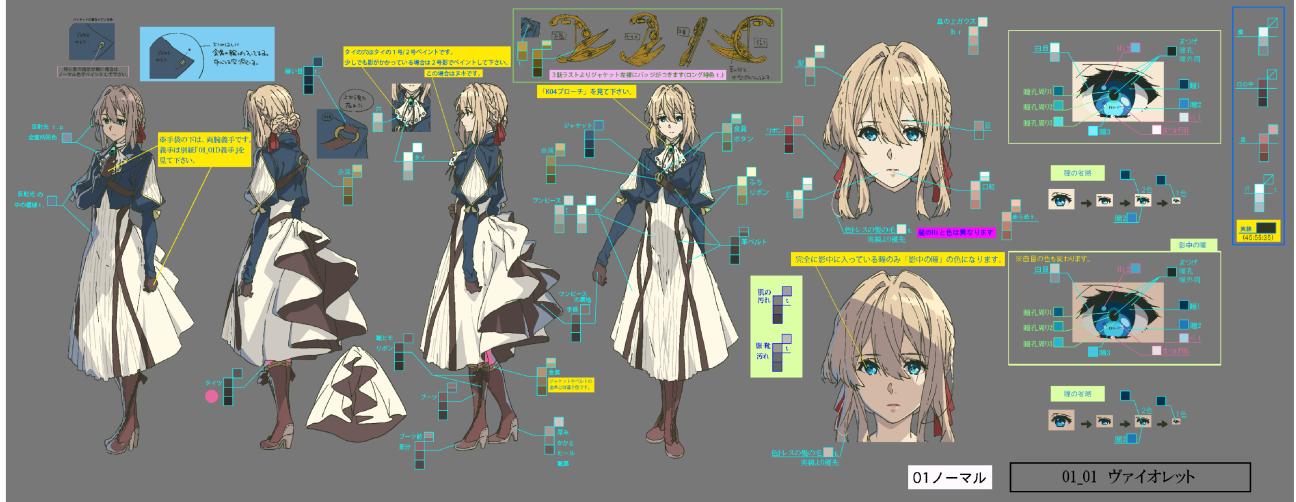


Fig. 6. Color chart for directing the colorization. Example from *Violet Evergarden* (2018).

simultaneously on distinct elements of the animation.

- **Layout (L/O):** Layouts (L/O) specify the placement and relationships between characters, objects, and backgrounds within a shot, based on storyboards. Artists ensure precise framing, perspective, and camera angles during this step. Layout approval marks the transition into more detailed work streams.
- **The 1st Keyframe Animation (Genga/1st Key):** The First keyframe animation (Genga/1st Key) provides the structural foundation for character movement and acting. Key animators produce rough sketches representing critical moments within a sequence, which serve as references for further refinement.
- **The 2nd Keyframe Animation (Nigen/2nd Key):** The Second keyframe animation (Nigen/2nd Key) builds upon Genga, adding detail to poses and smoothing out motion. This step bridges gaps in movement and ensures continuity. Corrections by animation directors are often incorporated at this stage. After this step, the resulting line drawing highlights are represented in red, while shadow lines are depicted in blue, as shown in the second

row of Figure 4.

- **Inbetweening:** Inbetweening involves drawing transitional frames between keyframes to achieve smooth and natural motion. This stage, typically handled by assistant animators, forms the bulk of traditional Cel-Animation and is critical for maintaining flow and consistency.²
- **Colorization:** Once line work is finalized, frames are colored based on established color models. Colorization not only enhances visual appeal but also reinforces the emotional and thematic tone of the animation.³ Teams often use digital tools for efficient and precise coloring.

- 3) **Post-production:** Post-production integrates all animated elements into a cohesive final product, adding effects, sound, and other finishing touches.

- **Compositing & Photography:** Compositing combines characters, backgrounds, and visual effects into finalized scenes. Traditionally involving photography of physical cel layers, modern workflows rely on digital compositing software to achieve similar results.

²<https://tips.clip-studio.com/en-us/articles/954>

³<https://setteidreams.net/color-designs/>

- **Cutting (CT):** Cutting organizes animated sequences into a narrative flow, ensuring proper timing, pacing, and synchronization with sound. Editors fine-tune scenes for coherence and alignment with the director's vision.
- **Music & Sound Effects:** Sound design includes creating and syncing background music (BGM) and sound effects (SE) to the visuals. These elements enhance emotional depth and provide a more immersive viewing experience.
- **After Recording (AR) & Dubbing (DB):** The dubbing phase involves recording and synchronizing voice acting with the completed animation. Additionally, this step integrates character performances with the visual flow of the animation.

4) Other Processes:

- **Background Design:** Background design relies on the BG Setting obtained during the Setting stage and Mood Board, as well as the L/O from the Layout stage. In addition to manual drawing, 3D assistance is sometimes required.
- **3D Assistance:** Place buildings and characters in 3D space to recreate the actual scene, and position a virtual camera within the 3D space to finalize the layout. The layout created in 3D is printed onto paper, and from there, the process of adding character movements by hand continues.
- **Clean-Up:** Clean-up is the process of refining rough sketches, aiming to transform the initial, hasty lines into clear and polished final line art. This step ensures the accuracy of details in characters and scenes and is typically carried out during the 2nd Keyframe Animation.
- **Quality Control & Inspection:** The animation production pipeline incorporates multiple inspection stages to ensure consistent quality. The details can be found in Appendix B.

B. Generative AI

1) *Large Language Models (LLMs):* Language models learn the joint probability $p(x_{1:L})$ over a token sequence $x_{1:L}$ via the chain rule:

$$p(x_{1:L}) = \prod_{i=1}^L p(x_i | x_{1:i-1}), \quad (1)$$

where L is the sequence length. With billions of parameters, LLMs utilize tokenizers and self-attention layers to predict token probabilities autoregressive: $\mathcal{M}(x_{1:i-1}) = p(x_i | x_{1:i-1})$. Decoding strategies, such as greedy decoding:

$$x_t = \arg \max_{s \in S} \log p_{\mathcal{M}}(s | x_{1:t-1}), \quad (2)$$

control token selection, while sampling strategies promote diversity.

Key characteristics of LLMs include:

- **Scaling Laws [36]:** Model performance grows predictably with increased parameters, data, and compute.
- **Emergent Abilities [37]:** At scale, models exhibit novel behaviors like in-context learning, instruction following, and chain-of-thought reasoning.

LLMs have been applied in animation and filmmaking to streamline tasks like scriptwriting, character backstory creation, and generating visual descriptions for storyboarding. Multimodal LLMs (MLLMs or LMMs) [9, 38] extend LLMs by integrating visual encoders and cross-modal aligners, excelling in tasks that require both visual and textual reasoning.

2) *GAN, VAE, and Diffusion Models:* Generative Adversarial Networks (GANs) [39] involve a generator G and a discriminator D in a minimax game:

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log (1 - D(G(z)))] \quad (3)$$

Applications include background generation and style transfer. StyleGAN [40] refines stylistic consistency, while Pix2Pix [41] transforms sketches into images.

Variational Autoencoders (VAEs) [42] model latent distributions by maximizing the Evidence Lower Bound (ELBO):

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - D_{KL}(q_{\phi}(z|x) \| p(z)), \quad (4)$$

where D_{KL} is the Kullback-Leibler divergence. VAEs excel in generating controllable character poses and expressions.

Diffusion models [10, 43] iteratively denoise samples drawn from Gaussian noise to approximate data distributions. A forward process adds noise in steps:

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I}), \quad (5)$$

and a neural network $\epsilon_{\theta}(x_t, t)$ predicts noise in the reverse process. The model minimizes the following Mean Square Error (MSE) loss:

$$\mathcal{L} = \mathbb{E}_{x_0, \epsilon, t} [\|\epsilon - \epsilon_{\theta}(x_t, t)\|^2], \quad (6)$$

where $x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon$, with $\alpha_t = \prod_{i=1}^t (1 - \beta_i)$. Models like Stable Diffusion [44] and ControlNet [12] extend this framework to tasks such as animation and colorization. Video-specific diffusion models address temporal consistency, further bridging key production gaps.

III. GENAI FOR CEL-ANIMATION

In this section, we explore how GenAI approaches are designed for each specific process in Cel-Animation, alongside those that show promising potential for Cel-Animation production despite their original development for other applications. We also collect related datasets, as shown in Table I.

A. GenAI for Pre-production

1) *Script Generation:* The script traditionally relies on human writers' creativity and narrative skills. However, the application of generative AI methods varies between original animations and adaptations. LLMs can be primarily used for original animations to generate story settings, expand details, and create dialogues from scratch. In contrast, for adaptations, such as those based on novels, manga, or games, LLMs first need to understand the source material's plot, character relationships, and narrative style before creating or adapting the script. With the advancements in proprietary (e.g., ChatGPT [5], Claude [47], Gemini [46]) and open-source pre-trained LLMs (e.g., LLaMA [1], Qwen [49]), generative

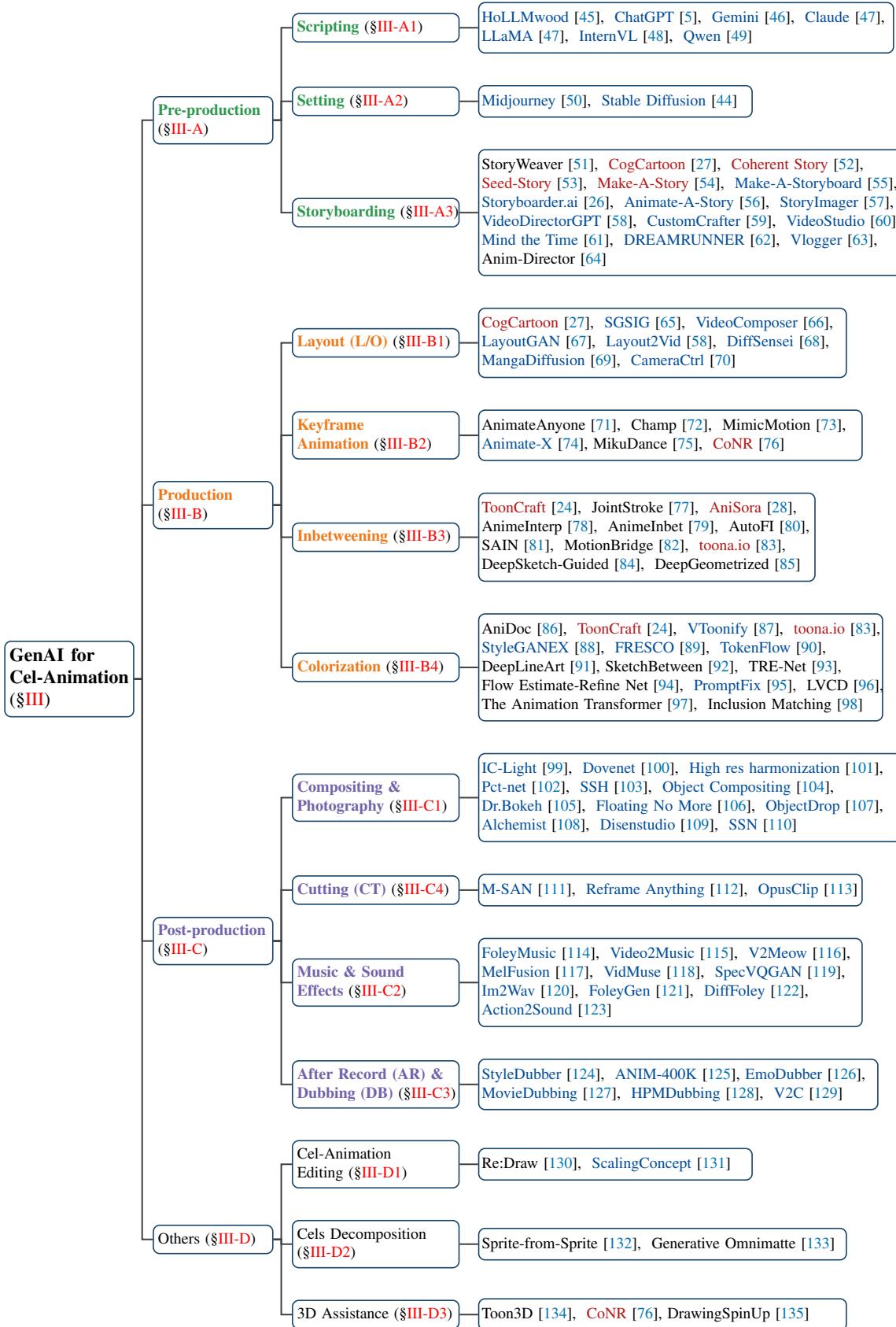


Fig. 7. The taxonomy of GenAI for Cel-Animation is primarily organized by the tasks involved in the production workflow. Key steps are highlighted by Green, Orange, and Purple. Methods capable of addressing multiple tasks within the Cel-Animation production process are indicated by Red, while Blue signifies methods that are not originally developed for Cel-related tasks but have the potential to be applied to Cel-Animation. In the future, there will be more and more models with non-blue colored fonts appearing in this figure.

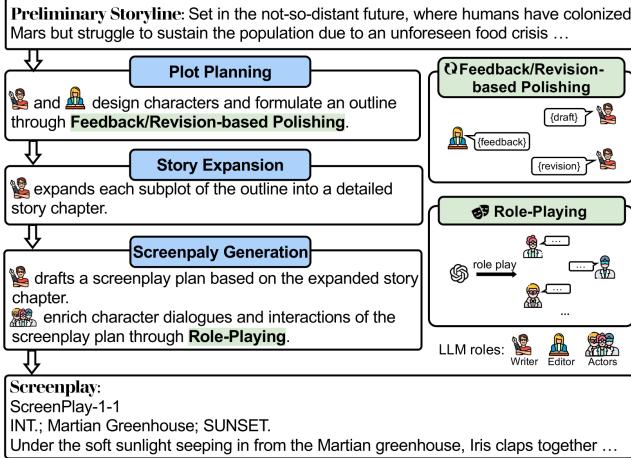


Fig. 8. LLMs for Script Generation: Users collaborate with generative AI to develop story scripts. The AI agent facilitates plot planning, story expansion, and screenplay creation through iterative feedback, revision-based refinement, and role-playing interactions. Example adapted from [45].

AI has demonstrated strong adaptability in both original and adapted scriptwriting, providing flexible and efficient solutions for animation script generation. The HoLLMwood [45] framework illustrated in Figure 8 introduces a role-playing mechanism, utilizing pre-trained LLMs to automate various stages of scriptwriting, including plot development and dialogue generation. This framework can generate structured scripts tailored to the requirements of different animation projects while optimizing character interactions and narrative logic.

2) **Setting Generation:** While there is no dedicated research focusing on character/object setting generation, many artists have already experimented with tools like Midjourney [50] and Stable Diffusion [44] to generate designs for items, characters, scenes, and mechas through text prompts. As shown in Figure 9, these generated contents demonstrate a certain degree of consistency (for example, multiple views of characters from different angles maintain good identity preservation). However, there are notable imperfections in the details. For instance, the generated character designs may include objects that are either absent from the character or lack practical significance; furthermore, the text appearing in these generated design sheets often consists of meaningless gibberish characters rather than coherent text.

3) **Storyboard Generation:** Storyboard generation primarily focusing on live-action film storyboards [58] while also addressing non-photorealistic storyboard generation such as animation [27, 51]. These systems typically accept textual scripts as input and generate visual storyboards in the form of sequential images or video sequences. To develop an automated storyboard production pipeline utilizing AI tools, users can provide high-level narrative guidelines to LLMs, which then generate detailed scene descriptions. These descriptions subsequently serve as input for image generation models, which produce the final sequential visual narrative in either static or animated form. The main challenge lies in achieving consistent identity in the generated images and ensure the

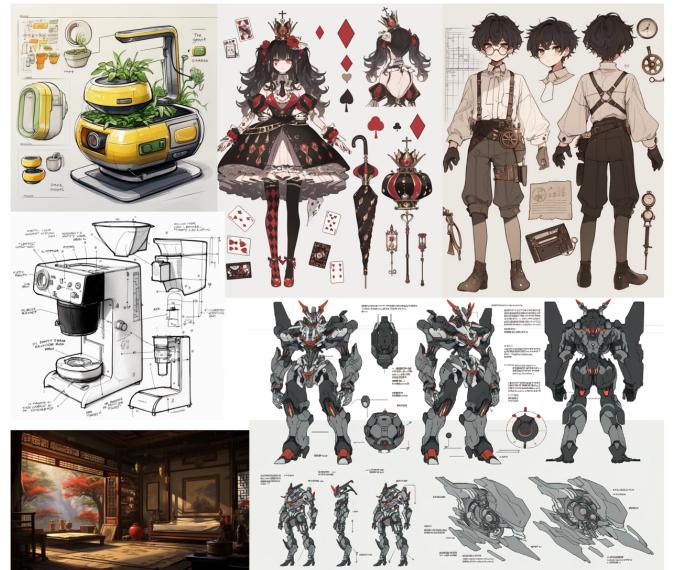


Fig. 9. Examples of AI-generated Settings with Midjourney [50] and Stable Diffusion [44], showing various designs including items, characters, scenes, and mecha.

alignment between human intent expressed in the script and the storyboard content, which determines whether the model can generate storyboards that accurately reflect human ideas. For identity preserving, recent works [52, 54] introduce memories that save the existing generation representation features or leverage attention controls [53]. Other works also enhance the text semantic correspondence between instruction and generation [51, 136]. CogCartoon [27] has introduced precise control over character positions within the storyboard frames, offering more flexibility compared to pure text-guided generation.

B. GenAI for Production

1) **Layout Generation:** During Cel-Animation production, storyboards are typically rough sketches that require a layout phase to further refine the positions and relationships of characters, scenes, and objects, as well as to elaborate on perspective, camera angles, and lighting information. While the previously discussed storyboard generation methods using bounding boxes can achieve a certain degree of position control, they cannot achieve finer-grained control over elements such as character poses and object orientations. This limitation is also common in many layout generation models [67]. However, as shown in the top of Figure 11, this layout generation approach [65] can achieve further refinement by accepting rough sketches as input and outputting layouts in the form of realistic images. Although rough, these sketches can effectively convey information about object size, pose, orientation, distance from camera, and perspective. Nevertheless, this method primarily focuses on object layout, and may perform less effectively for character layouts, particularly as it mainly supports photorealistic output generation, which could compromise anime character inputs. Furthermore, there is a lack of video-oriented layout capabilities, as current research remains predominantly focused on layout generation

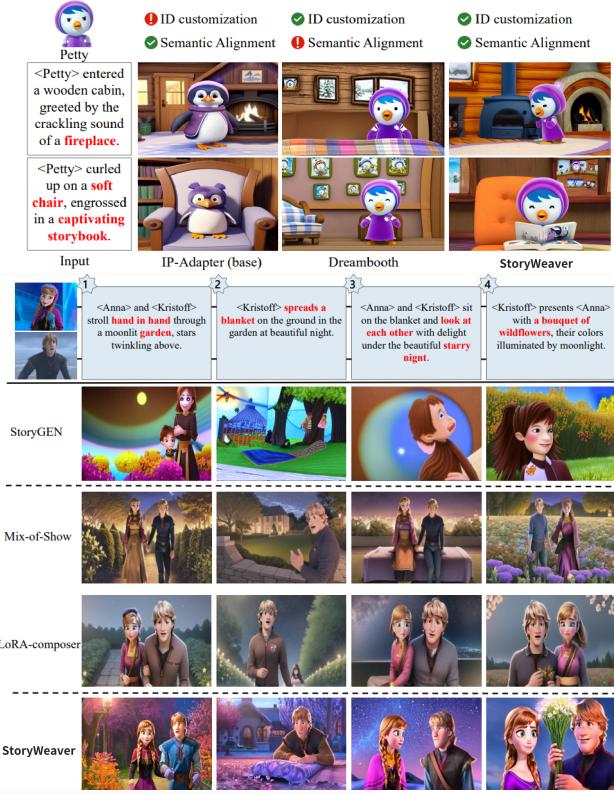


Fig. 10. AI-generated Storyboard. Identity preservation during sequential generation and alignment between story scripts and image semantics are the two core components of storyboard generation. This figure illustrates an example adapted from StoryWeaver [51].

in static images. Video generation models with camera motion control [70] enable simultaneous foreground and background motion, addressing the limitations of static layout models, as shown in the bottom of Figure 11. These approaches are useful for Cel-Animation, offering dynamic layouts that support various camera angles, distances, and continuity across frames, bridging gaps in traditional static layout methods.

2) **Keyframe Animation:** The keyframe animation mainly focuses on the moving elements within a scene. Therefore, we mainly collect and review methods focused on character animation generation [71]. To achieve this goal, recent works [137] leverage 2D body landmarks for the target avatar pose control, which provides explicit spatial guidance for human pose generation. For example, Animate-Any-One [71], Champ [72], and MusePose [138] introduce ReferenceNet to incorporate the spatial body information for the target pose generation, effectively bridging the gap between source and target poses. These approaches demonstrate significant improvements in pose transfer accuracy and natural motion synthesis. MimicMotion [73] and Animate-X [74] further advance this direction by considering the different body sizes and proportions to achieve better generalization across realistic persons and cartoon characters. This consideration of body morphology variations enables more robust pose transfer between subjects with different physical characteristics, expanding the applicability of these methods to a broader range

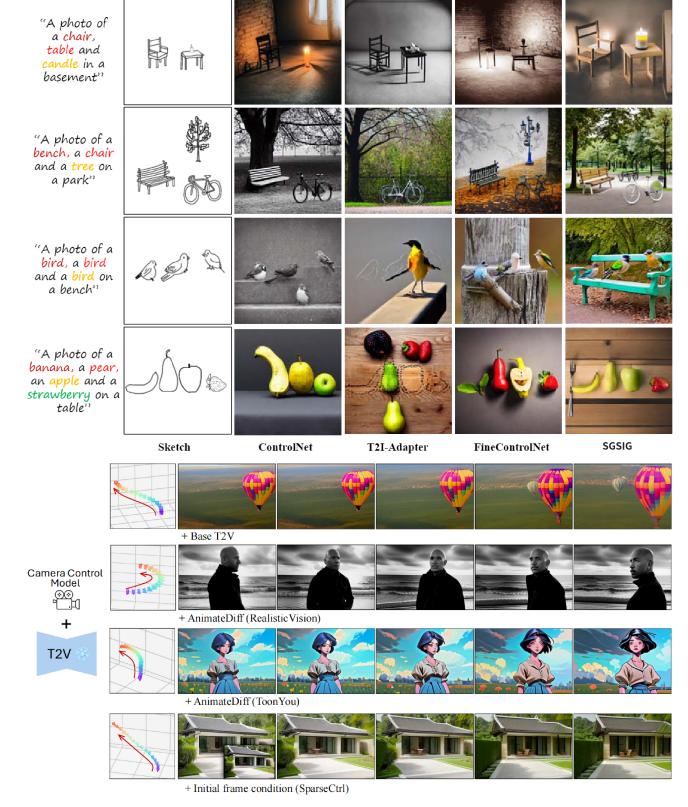


Fig. 11. AI-generated Layouts. **(Top)** Layout generation guided by sketches and textual descriptions using various methods, showcasing their ability to refine object placement and appearance. **(Bottom)** Camera motion control enables synchronized foreground and background motion, dynamic perspective shifts, and continuous scene adjustments across frames. Adapted from [65] and [70].

of scenarios. These methods aim to regulate the movement of the body, while there are also many methods that focus on the movement control of the face. Together, they form a comprehensive framework for character animation. Recent works [29, 139]–[141] animate the photo-realistic/toonification face from 3D Gaussian splatting or neural representation. For example, TextToon [29] and Emo-Avatar [141] adopt the 3D Gaussian representation with LLMs, and enabling the adaptation of facial cartoon style through text-based modifications. Ada-TalkingHead [139] employs neural keypoints to drive head animation, while Editable-Head [140] generates head animations using explicit landmark-based representations. In contrast to coarse body control, these methods are more refined, offering greater sophistication and intricate details for facial motion animation.

3) **Inbetweening:** Inbetweening is one of the most mechanically straightforward yet labor-intensive and time-consuming steps in the animation pipeline. This makes it particularly suitable for efficiency enhancement through GenAI. Consequently, compared to other steps in the animation production process, there has been a relatively substantial of research in AI for inbetweening. The current mainstream approach, such as Tooncraft [24], AutoFI [80], etc., draws inspiration from video interpolation models, where diffusion models are employed to predict intermediate frames given the first and last frames.

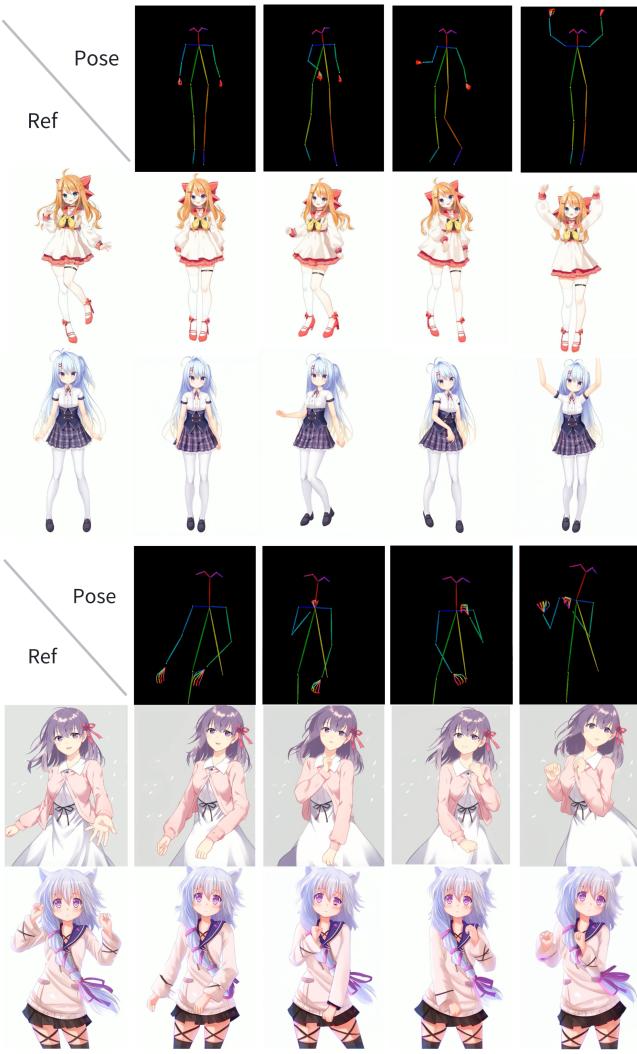


Fig. 12. Keyframe animation with AI. Conditioned on the source image and the target 2D landmarks, the model is capable of generating the deformed avatar with the target pose. Example adapted from [71].

[84] introduces a sketch-driven deep learning framework to automatically generate smooth and stylistically consistent in-between frames for cartoon animation. By leveraging user-provided sketches between key frames, the method captures desired motions and deformations while preserving character details and visual style. SAIN [81] utilizes a multi-stream U-Transformer with self and cross-attention based building blocks to capture region, stroke, and pixel-level guidance of key frames, producing inbetween frames with smoother transitions and finer details. [85] geometrizes cartoon line arts as geometrized vector graphs, converting the inbetweening task as a graph fusion problem with vertex repositioning, and proposes AnimeInbet, a deep learning-based framework, to capture the sparsity and unique structure of line drawings while preserving the details during inbetweening. [79] introduces an “inbetween chart” that visually maps out potential movement trajectories for each articulating part. By leveraging trajectory-guided sliders, animators can make intuitive, fine-grained adjustments to object transformations, rotations, and deformations. [77] introduces a method to construct inbe-

tween frames by simultaneously recovering high-level stroke structures from 2D animation frames and establishing correspondence across them. [82] proposes MotionBridge which employs a learnable, structured representation of temporal and spatial cues, ensuring both temporal coherence and enabling users to control over the details of intermediate frames.

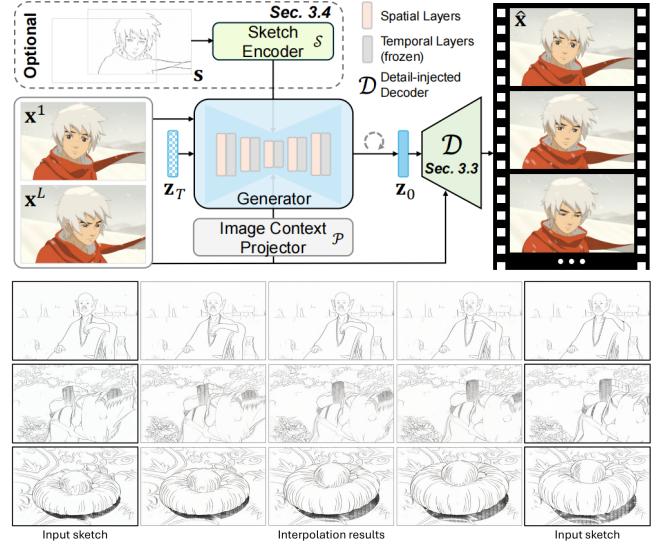


Fig. 13. GenAI for inbetweening and interpolation. The model can generate inbetrwens with sketches and colored frames as input. Example adapted from [24].

4) **Colorization:** Colorization is another labor-intensive step that has been extensively studied. Since our survey primarily focuses on Cel-Animation, we have concentrated our review on anime video colorization rather than static anime image colorization. The input typically consists of line drawings along with a colored line drawing as guidance, enabling the colorization of all black-and-white line drawings [23]. More recently, this transformation has been approached as a style transfer problem in computer vision. TextToon [29], StyleGANEX [88] and VToonify [87] simplifying the process through few-shot learning, where models learn domain transfer from source sketches to target styles using only a small set of example animations. However, these approaches often struggle with maintaining temporal consistency across video sequences and only limited to avatars. The latest researches leveraging large-scale pretrained Diffusion Models [89, 90, 142] has achieved video-level style transfer even in zero-shot scenarios, significantly reducing manual labor requirements. Recent works [94, 96]–[98] are capably of achieving more complex colorizations for the whole scene with precise control. Notably, some current models can simultaneously perform both inbetweening and colorization steps, requiring only one colored line drawing as reference along with first and last frames as input to generate intermediate frames that are already colored. For example, ToonCraft [24] demonstrates this capability by supporting both inbetweening and colorization in a single model.

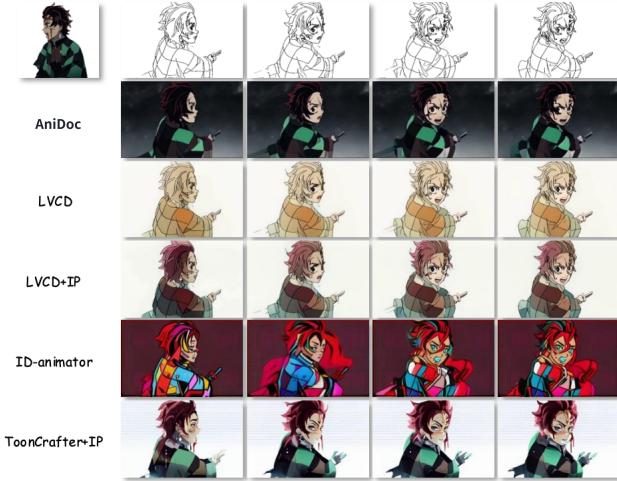


Fig. 14. Colorization. Stylization models achieve domain transfer from sketch images to target picture styles. This example is adapted from [86].

C. GenAI for Post-production

1) **Compositing & Photography:** After animation completes inbetweening and colorization, it needs to be composited with backgrounds and other materials. Compositing involves more than simply overlaying foreground elements onto backgrounds; it requires careful attention to maintaining consistency between foreground and background elements. Although the spatial relationships and perspective between foreground and background are unified during the layout phase, specific adjustments are still necessary. For instance, in twilight scenes, the environmental lighting and color tones of foreground and background elements need to be harmonized; the variations in light and shadow, as well as the reflection of light on the character's body, are particularly challenging to handle. In other cases, particle effects, special textures, or flame effects may need to be added. All these processing steps must be completed during the Photography phase. Considerable researches [100]–[104, 106] have already explored methods for harmonizing foreground and background composition (not limited to video). For example, [143] attempted to adjust objects when compositing them into backgrounds to better integrate them into the environment, while [99] explored controllable lighting editing, as shown in Figure 15.

2) **Music & Sound Effects:** Audio plays an essential role in animation production, creating an immersive experience and engaging the audience in the story being told. Specifically, audio can be categorized into three primary sources: speech, music, and sound effects. Among these, speech, commonly pre-recorded, undergoes integration during the dubbing process. This section focuses on the production of music and sound effects.

- **Music Generation:** Music in animation often serves to amplify character emotions and enhance the atmosphere. For instance, an energetic song can heighten the intensity of a battlefield scene, while a melancholic tune can evoke empathy and draw the audience into a poignant narrative. Recently, AI-generated music has emerged as



Fig. 15. Compositing & Photography: This example adapted from [99] demonstrate the capability of achieving illumination harmonization and editing with Diffusion Models.

a significant area of innovation, attracting considerable attention for its potential to revolutionize audio production in animation. A prominent approach in this domain involves generating music from text descriptions, leveraging techniques such as diffusion models [144]–[146] or sequence-to-sequence modeling [147]–[149]. However, text-guided music generation often proves suboptimal for video, where synchronization between visual content and generated music is critical. This limitation has spurred advancements in video-guided audio generation.

Relevant to this survey is the development of video-guided music generation methods, including FoleyMusic [114], Video2Music [115], V2Meow [116], MeFusion [117], and VidMuse [118]. While these methods are not specifically tailored for anime music generation, they offer valuable, generalizable tools that lay the foundation for future research in AI-driven animation.

- **Sound Effects Generation:** Sound effects encompass a wide range of non-speech and non-music sounds from daily life, such as animal noises, traffic sounds, clanging or hitting effects and more. These sounds are typically specific to the scenes depicted in the video, presenting significant challenges in video-audio alignment. Specifically, generating sound effects that match the semantic content of the scene and synchronizing them with precise timing is crucial to achieve harmony between the audio and visual elements. Several approaches have addressed this challenge using auto-regressive transformer models to generate audio from visual features, including methods such as SpecVQGAN [119], Im2Wav [120], and FoleyGen [121]. Diffusion models have also gained popularity in this domain [122, 123, 150].

By leveraging automatic tools for generating audio synchronized with video content, AI-driven animation holds the

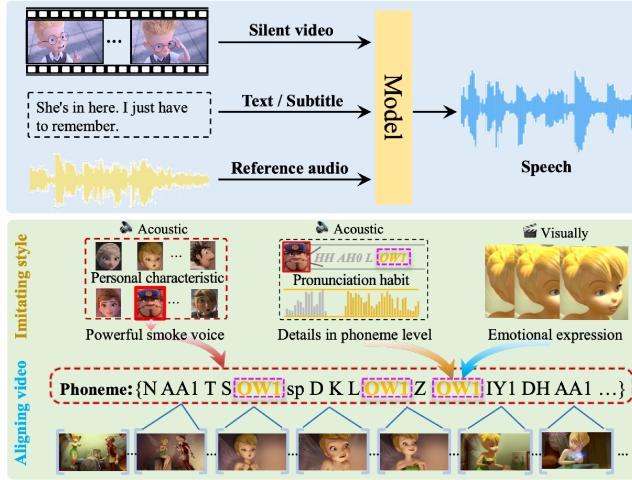


Fig. 16. GenAI for Dubbing. During this procedure, models detect the facial regions of the target avatar and replace with synchronized facial expressions and lip movements conditioned on given audio. The figure depicts an example sourced from [124].

potential to significantly reduce human labor while enriching the audio-visual experience.

3) **Dubbing (DB)**: In animation production, dubbing encompasses the broader process of ensuring synchronization between audio and visual elements, particularly focusing on the correspondence between speech and character movements. Traditionally, this process follows two main approaches: after-recording, where the animation is adjusted to match pre-recorded voice acting, and dubbing, which involves post-production audio processing to synchronize with existing animations⁴. Recent AI-powered dubbing models like EmotiVoice [124] have demonstrated the potential to automate this synchronization process. These systems employ a comprehensive approach: first detecting and isolating the character's facial region, then analyzing the target audio to guide the deformation or regeneration of facial features, particularly mouth movements and expressions. The process involves analyzing silent video, text/subtitles, and reference audio to generate emotionally expressive speech that aligns with the visual elements. These models achieve natural synchronization by combining acoustic characteristics from reference voices, pronunciation patterns from phoneme-level analysis, and emotional expressions derived from visual cues, enabling the generation of synchronized and emotionally appropriate speech-motion correspondence for animated characters.

4) **Cutting**: Cutting refers to the process of arranging shots produced during the photography phase according to the storyboard script, and adjusting the duration, rhythm, and sequence of shots based on directorial intent to ensure proper pacing and visual presentation while conforming to television format and runtime requirements. Currently, there is no GenAI work specifically focused on animation cutting. As shown in

⁴In this survey, we use the term “dubbing” to refer to the general pipeline of maintaining consistency between animation (video) and audio/speech, encompassing both after-recording and traditional dubbing approaches in animation production.

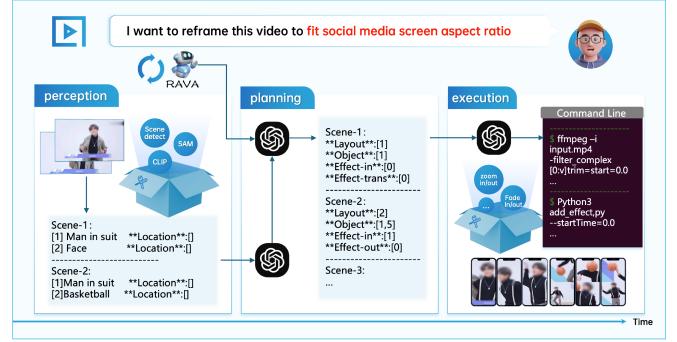


Fig. 17. GenAI-enabled video cutting workflow using RAVA [112], which can cut video by leveraging scene detection, planning, and automated FFmpeg execution.

Figure 17, Reframe Anything (RAVA) [112] employs LLM to generate FFmpeg instructions to reframe and restructure video content for different aspect ratios while preserving the most salient elements, making it a versatile tool in visual media production. M-SAN [111] concentrates on advertisement video cutting, taking a long video and specified duration as input, and is tasked with extracting multiple segments to create a shorter video that preserves important content from the original, maintains content coherence, and meets the given duration requirements. Despite theirs current focus, these approaches show potential for expansion into the Cel-Animation domain.

D. Others

1) **Cel-Animation Editing**: Recent advances in context-aware translation have enabled more precise and controllable editing of cel-animations. Re:Draw [130] proposes a novel framework that combines the benefits of inpainting and image-to-image translation while respecting both original content and contextual relevance. This approach allows for local edits while maintaining global consistency, particularly useful for enhancing details like character eyes without compromising the overall animation style. The method employs dual discriminator structures and novel adversarial losses to ensure both artistic control and contextual coherence, demonstrating significant improvements over traditional editing approaches in preserving animation consistency. ScalingConcept [131] introduces a training-free editing method for images using pre-trained text-guided diffusion models. Among its diverse applications, one is specifically tailored for anime editing. During the photography and post-production stages of anime production, cumulative errors in line processing can lead to blurred lines, causing the image to appear fuzzy. This issue is often exacerbated by filters used in scenes like sunsets, which cannot be resolved by merely increasing the resolution or bitrate of the anime. With the ScalingConcept method, such issues can be addressed, resulting in a significant improvement in the overall visual quality of the anime.

2) **Cels Decomposition**: Decomposing animation into individual layers or “sprites” is crucial for both analysis and editing. Sprite-from-Sprite [132] introduces a self-supervised framework that can automatically decompose cartoon ani-

TABLE I

COMPREHENSIVE OVERVIEW OF DATASETS FOR CEL-ANIMATION AND RELATED TASKS: THIS TABLE PROVIDES A DETAILED SUMMARY OF DATASETS EXPLICITLY DESIGNED FOR CEL-ANIMATION TASKS—SUCH AS SAKUGA-42M, ANISORA, ANIMERUN, ATD-12K, MIXAMO-LINE240, ANT, AND PAINTBUCKET-CHARACTER—FOCUSING ON OBJECTIVES LIKE ANIMATION COMPREHENSION, GENERATION, INBETWEENING, AND COLORIZATION. IT ALSO HIGHLIGHTS DATASETS NOT SPECIFICALLY CREATED FOR CEL-ANIMATION BUT HIGHLY RELEVANT TO ADJACENT TASKS, INCLUDING ANIM-400K, V2C-ANIMATION, iHARMONY4, SSHARMONIZATION, HIVIDIT, ALCHEMIST, MANGAZERO, AND ADS-1K, WHICH SUPPORT APPLICATIONS LIKE DUBBING, HARMONIZATION, LAYOUT, AND SCENE CUTTING. KEY METRICS, SUCH AS THE NUMBER OF CLIPS AND SAMPLES, ARE INCLUDED TO EMPHASIZE THE SCOPE, SCALE, AND APPLICABILITY OF THESE RESOURCES IN ANIMATION RESEARCH AND DEVELOPMENT.

Dataset	Format	Task in Cel	#Clips	#Samples
Sakuga-42M [151]	Video	Comprehension, Generation	1.2M	42M
AniSora [28]	Video	Comprehension, Generation	10M	
Anim-400K [125]	Video	Dubbing		400K
V2C-Animation [129]	Video	Dubbing	10,217	
iHarmony4 [100]	Image	Compositing, Photography (Harmonization)		73,146
SSHarmoziation [103]	Image	Compositing, Photography (Harmonization)		216
HVIDIT [152]	Image	Compositing, Photography (Harmonization)		3,336
Alchemist [108]	Image	Compositing, Photography (T2I Generation)		
AnimeRun [153]	Video	Inbetweening, Colorization	30	2,891
ATD-12K [78]	Image	Inbetweening		12,000 triplets
Mixamo-Line240 [85]	Image	Inbetweening		49,262
Ant [97]	Image	Colorization		11,000
PaintBucket-Character [98]	Image	Colorization		14,545
MangaZero [68]	Image	Layout		470K
Manga109Story [69]	Image	Layout		21K
Ads-1k [111]	Video	Cutting	1K	
ANITA DATASET [154]	Image	Genga, Nigen, Colorization		16,000

mations into reusable sprite elements. The method recognizes that sprites in real-world cartoons can vary between simple, computationally-tractable animations and complex, hand-drawn sequences. Similarly, Generative Omnimatte [133] advances this concept by incorporating a generative video prior to handle dynamic backgrounds and complete occluded regions, enabling more robust decomposition of animation elements while preserving associated effects like shadows and reflections.

3) *3D Assistance*: Three-dimensional assistance tools have emerged to support traditional cel-animation production. DrawingSpinUp [135] introduces a method for automatically animating character drawings in 3D space, enabling more complex movements while maintaining artistic integrity. Likewise, Toon3D [134] focuses on reconstructing 3D scenes from cartoon drawings, allowing for novel viewpoint generation and better planning of complex animation sequences. These approaches bridge the gap between 2D and 3D animation techniques, providing artists with tools to create more dynamic sequences while preserving the distinctive qualities of hand-drawn animation.

IV. DISCUSSION

A. Limitations

Despite the advancements and potential applications of GenAIs in Cel-Animation, several limitations hinder their effectiveness and adoption.

1) *Expressiveness and Prompt Limitations*: The reliance on natural language prompts to direct generative models [60, 66] often fails to encapsulate the nuances of Cel-Animation. Elements such as exaggerated expressions, dynamic perspectives, and intricate actions are difficult to describe effectively

in textual form, limiting the scope of creativity achievable through prompt-based generation.

2) *Insufficient Reference Utilization*: Current GenAIs struggle to incorporate references like color charts and settings of items, characters, or scenes. This limitation results in outputs that lack fidelity to established design elements, creating inconsistencies that require additional manual intervention.

3) *Background Integration and Input Deviations*: Current models for tasks like colorization and inbetweening produce results with backgrounds, which complicates post-generation editing. Additionally, the input sketches for colorization or inbetweening are frequently derived from models rather than original keyframe animation.

4) *Limited Control and Artistic Freedom*: While automation provides efficiency, it limits the granular control artists can exercise over the final product. Full reliance on prompt-based systems undermines the creative autonomy that animators value, as these systems fail to allow detailed adjustments at varying levels of granularity.

5) *Legal and Copyright Concerns*: The use of Cel-Animations for training may raise copyright issues, which pose ethical and legal challenges for widespread adoption.

B. Future Directions

This section examines future directions and highlights key areas that warrant further improvement. As noted earlier, contemporary GenAI methods have been successfully applied to pre-production, production, and post-production processes and have achieved promising results. However, these approaches have yet to address all of the fundamental challenges inherent in cel-animation. In this study, we discuss four key areas that can be improved by data curation and model design.

1) Building Large-Scale and Comprehensive Celluloid Animation Dataset: Datasets play an important role in building robust celluloid animation production models. Although some datasets are available [85, 98, 151, 153], they remain limited for several critical processes, leaving gaps that hinder further advancement in this domain. For instance, script generation frequently relies on text corpora derived from fiction and cartoons, but such resources rarely capture the nuanced interplay between visual storytelling and narrative structure unique to hand-drawn animation. As a result, many GenAI models struggle to generate consistent, contextually rich scripts that can be directly integrated into a production pipeline. A promising way to address these limitations is to develop new large-scale, comprehensive datasets that mirror real-world animation workflows. Such datasets should not merely expand in scale but also diversify in content to cover a broad range of animation styles, genres, and production stages. For example, by incorporating both preliminary sketches and final colored frames—along with corresponding scripts or dialogue—researchers can more effectively train models to understand the full spectrum of visual and narrative elements that define hand-drawn animation.

2) Advanced LLM-based Content Generation Methods: While there have been numerous efforts to use LLMs for script generation, existing approaches have limitations in the use of LLMs and have not incorporated multimodal content understanding. Recent advancements in MLLMs [5, 7, 8, 38, 155] have demonstrated exceptional performance in vision-language tasks [156]–[160], underscoring the necessity of holistic, multi-phase methods that seamlessly integrate textual and visual elements. Such approaches will facilitate Cel-Animation with MLLMs’ powerful multimodal understanding and reasoning capabilities. A key challenge lies in aligning textual narrative with the fine-grained visual components that define Cel-Animation, such as character expressions, background design, and color schemes. Existing methods often treat these tasks independently (e.g., generating a script first and later producing accompanying images) leading to potential mismatches in style, tone, or narrative context. Instead, an advanced LLM-based approach could employ multimodal reasoning, where scene descriptions and dialogue generation run in tandem with storyboard sketches, character poses, and layout proposals. By iteratively refining both text and images together, the framework ensures that each creative decision — ranging from dialogue shifts to lighting changes — is consistently reflected across the entire production pipeline.

3) Unified Multiprocess Generation Frameworks: Ensuring coherence in model-generated animation content is a fundamental requirement. Certain processes in the celluloid animation production workflow can be unified within a single model, thereby enhancing consistency across each stage. Recent advances in autoregressive multimodal models [161]–[164] have further enabled language models to generate interleaved text and images, paving the way for more streamlined production pipelines. By producing storyline elements alongside corresponding visuals in a single pass [165]–[167], these models can streamline traditionally separate tasks such as script generation, storyboard layout, and character design.

For instance, while generating a dialogue, the same model can simultaneously propose camera angles or background sketches, ensuring textual and visual narratives evolve together rather than in isolation. These interleaved frameworks pave the way for more cohesive production pipelines: scene descriptions, character references, and camera instructions can be developed in tandem, eliminating the need for manual adjustments whenever upstream changes occur. As a result, both narrative and visual consistency are maintained, making it easier to implement revisions and reducing the likelihood of inconsistencies creeping into the final output. Ultimately, such unified multiprocess generation frameworks have the potential to significantly streamline celluloid animation production, bringing smoother handoffs and more coherent content to each stage of the pipeline.

4) Multi-Agent System for Script-to-Animation Generation: Since the production of celluloid animation is both complex and labor-intensive, it is challenging to address the entire process in an end-to-end manner. A feasible approach is to develop multi-agent systems that incorporate various expert models, enabling the simulation of each stage in the celluloid animation production workflow. Recent work on MLLM-based multi-agent systems [168]–[173] has shown promising results in tackling complex real-world tasks [174]–[176]. It is natural to employ expert models to each process in the workflow. Specifically, one could assign a script-generation agent to craft coherent narratives, a layout agent to refine camera angles and scene compositions, a keyframe animation agent to produce stable character motions, and an inbetweening agent to fill in smooth transitions between keyframes. Colorization and compositing agents would ensure that visual consistency is maintained across scenes and properly integrated with backgrounds and effects, while a dubbing agent would handle speech synchronization and emotional nuance. By orchestrating these task-specific agents within a cohesive multi-agent framework, it becomes possible to manage the interdependencies among different stages, promptly revise outputs when upstream changes occur, and ultimately deliver high-quality results that retain the unique charm of celluloid animation.

V. CONCLUSION

This survey explores the transformative impact of GenAI on traditional Cel-Animation workflows. It traces the evolution of Cel-Animation from handcrafted techniques to computer-assisted methods and now to the GenAI-driven era. GenAI methods such as diffusion models and LMMs are being utilized across all stages of production, from pre-production to post-production, significantly improving efficiency, accessibility, and creative opportunities while reducing repetitive tasks and high costs. However, there are many challenges remaining, including ensuring stylistic consistency, visual coherence, and balancing AI-generated content with human creativity. This survey concludes by emphasizing the potential of further innovation through curated datasets, advanced video generation methods, and unified multiprocess frameworks, underscoring the future role of GenAI in revolutionizing Cel-Animation.

REFERENCES

- [1] H. Touvron, T. Lavigil, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [2] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, “Qwen2. 5 technical report,” *arXiv preprint arXiv:2412.15115*, 2024.
- [3] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Lucchini, F. Yvon, M. Gallé *et al.*, “Bloom: A 176b-parameter open-access multilingual language model,” 2023.
- [4] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *Advances in neural information processing systems*, vol. 36, 2024.
- [5] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [6] Y. Tang, J. Bi, S. Xu, L. Song, S. Liang, T. Wang, D. Zhang, J. An, J. Lin, R. Zhu *et al.*, “Video understanding with large language models: A survey,” *arXiv preprint arXiv:2312.17432*, 2023.
- [7] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, “Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution,” *arXiv preprint arXiv:2409.12191*, 2024.
- [8] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma *et al.*, “How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites,” *arXiv preprint arXiv:2404.16821*, 2024.
- [9] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “Minigpt-4: Enhancing vision-language understanding with advanced large language models,” *arXiv preprint arXiv:2304.10592*, 2023.
- [10] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [11] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” *arXiv preprint arXiv:2010.02502*, 2020.
- [12] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.
- [13] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, “Scaling rectified flow transformers for high-resolution image synthesis,” in *Forty-first International Conference on Machine Learning*, 2024.
- [14] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, “Align your latents: High-resolution video synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22563–22575.
- [15] StudioBinder, “What is cel animation — examples, techniques & history,” *StudioBinder Blog*. [Online]. Available: <https://www.studiobinder.com/blog/what-is-cel-animation-definition/>
- [16] Wikipedia, “Computer animation production system.” [Online]. Available: https://en.wikipedia.org/wiki/Computer_Animation_Production_System
- [17] Toonz. [Online]. Available: <https://en.wikipedia.org/wiki/Toonz>
- [18] OpenToonz. [Online]. Available: <https://opentoonz.github.io/e/>
- [19] Wikipedia, “Adobe Flash.” [Online]. Available: https://en.wikipedia.org/wiki/Adobe_Flash
- [20] Celsys, “Clip Studio Paint.” [Online]. Available: <https://www.clipstudio.net/en/>
- [21] S. J. Napier, “Anime from akira to howl’s moving castle: Experiencing contemporary japanese animation,” *Palgrave Macmillan*, 2005.
- [22] J. Guajardo, O. Bursalioglu, and D. B. Goldman, “Generative ai for 2d character animation,” *arXiv preprint arXiv:2405.11098*, 2024.
- [23] Y. Meng, H. Ouyang, H. Wang, Q. Wang, W. Wang, K. L. Cheng, Z. Liu, Y. Shen, and H. Qu, “Anidoc: Animation creation made easier,” *arXiv preprint arXiv:2412.14173*, 2024.
- [24] J. Xing, H. Liu, M. Xia, Y. Zhang, X. Wang, Y. Shan, and T.-T. Wong, “Tooncrafter: Generative cartoon interpolation,” *arXiv preprint arXiv:2405.17933*, 2024.
- [25] A. Singh, “Future of animated narrative and the effects of ai on conventional animation techniques,” *7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, 2023.
- [26] Storyboarder.ai. [Online]. Available: <https://www.storyboarder.ai/>
- [27] Z. Zhu and J. Tang, “Cogcartoon: Towards practical story visualization,” *arXiv preprint arXiv:2312.10718*, 2023.
- [28] Y. Jiang, B. Xu, S. Yang, M. Yin, J. Liu, C. Xu, S. Wang, Y. Wu, B. Zhu, X. Zhang, X. Zheng, J. Xu, Y. Zhang, J. Hou, and H. Sun, “Anisora: Exploring the frontiers of animation video generation in the sora era,” *arXiv preprint arXiv:2412.10255*, 2024.
- [29] L. Song, L. Chen, C. Liu, P. Liu, and C. Xu, “Texttoon: Real-time text toonify head avatar from single video,” in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.
- [30] C. Li, D. Huang, Z. Lu, Y. Xiao, Q. Pei, and L. Bai, “A survey on long video generation: Challenges, methods, and prospects,” *arXiv preprint arXiv:2403.16407*, 2024.
- [31] W. Lei, J. Wang, F. Ma, G. Huang, and L. Liu, “A comprehensive survey on human video generation: Challenges, methods, and prospects,” *arXiv preprint arXiv:2407.08428*, 2024.
- [32] Z. Xing, Q. Feng, H. Chen *et al.*, “A survey on video diffusion models,” *arXiv preprint arXiv:2405.03150*, 2024.
- [33] P. Zhou, L. Wang, Z. Liu *et al.*, “A survey on generative ai and llm for video generation, understanding, and streaming,” *arXiv preprint arXiv:2404.16038*, 2024.
- [34] J. Cho, F. D. Puspitasari, S. Zheng *et al.*, “Sora as an agi world model? a complete survey on text-to-video generation,” *arXiv preprint arXiv:2403.05131*, 2024.
- [35] Y. Zhao, D. Ren, Y. Chen, W. Jia, R. Wang, and X. Liu, “Cartoon image processing: A survey,” *International Journal of Computer Vision*, vol. 130, no. 11, pp. 2733–2769, 2022.
- [36] J. Kaplan *et al.*, “Scaling laws for neural language models,” *arXiv preprint arXiv:2001.08361*, 2020.
- [37] W. Zhao *et al.*, “A survey of emergent abilities in large language models,” *arXiv preprint arXiv:2304.14200*, 2023.
- [38] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, “Visual chatgpt: Talking, drawing and editing with visual foundation models,” *arXiv preprint arXiv:2303.04671*, 2023.
- [39] I. Goodfellow *et al.*, “Generative adversarial nets,” *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680, 2014.
- [40] T. Karras *et al.*, “A style-based generator architecture for generative adversarial networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4217–4228, 2019.
- [41] P. Isola *et al.*, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.
- [42] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [43] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” *arXiv preprint arXiv:1503.03585*, 2015.
- [44] 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*, 2022, pp. 10684–10695.
- [45] J. Chen, X. Zhu, C. Yang, C. Shi, Y. Xi, Y. Zhang, J. Wang, J. Pu, R. Zhang, Y. Yang *et al.*, “Hollmwood: Unleashing the creativity of large language models in screenwriting via role playing,” *arXiv preprint arXiv:2406.11683*, 2024.
- [46] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, Alayrac *et al.*, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” *arXiv preprint arXiv:2403.05530*, 2024.
- [47] Anthropic, “Claude 3.5 sonnet news.” [Online]. Available: <https://www.anthropic.com/news/clause-3-5-sonnet>
- [48] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, “Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24185–24198.
- [49] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, “Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond,” *arXiv preprint arXiv:2308.12966*, 2023.
- [50] MidJourney Team, “MidJourney: Generative AI Art Tool,” 2024.
- [51] J. Zhang, J. Tang, R. Zhang, T. Lv, and X. Sun, “Storyweaver: A unified world model for knowledge-enhanced story character customization,” *arXiv preprint arXiv:2412.07375*, 2024.
- [52] X. Pan, P. Qin, Y. Li, H. Xue, and W. Chen, “Synthesizing coherent story with auto-regressive latent diffusion models,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 2920–2930.
- [53] S. Yang, Y. Ge, Y. Li, Y. Chen, Y. Ge, Y. Shan, and Y. Chen, “Seed-story: Multimodal long story generation with large language model,” *arXiv preprint arXiv:2407.08683*, 2024.

- [54] T. Rahman, H.-Y. Lee, J. Ren, S. Tulyakov, S. Mahajan, and L. Sigal, “Make-a-story: Visual memory conditioned consistent story generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2493–2502.
- [55] S. Su, L. Guo, L. Gao, H. T. Shen, and J. Song, “Make-a-storyboard: A general framework for storyboard with disentangled and merged control,” *arXiv preprint arXiv:2312.07549*, 2023.
- [56] Y. He, M. Xia, H. Chen, X. Cun, Y. Gong, J. Xing, Y. Zhang, X. Wang, C. Weng, Y. Shan *et al.*, “Animate-a-story: Storytelling with retrieval-augmented video generation,” *arXiv preprint arXiv:2307.06940*, 2023.
- [57] M. Tao, B.-K. Bao, H. Tang, Y. Wang, and C. Xu, “Storyimager: A unified and efficient framework for coherent story visualization and completion,” in *European Conference on Computer Vision*. Springer, 2025, pp. 479–495.
- [58] H. Lin, A. Zala, J. Cho, and M. Bansal, “Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning,” *arXiv preprint arXiv:2309.15091*, 2023.
- [59] T. Wu, Y. Zhang, X. Wang, X. Zhou, G. Zheng, Z. Qi, Y. Shan, and X. Li, “Customcrafter: Customized video generation with preserving motion and concept composition abilities,” *arXiv preprint arXiv:2408.13239*, 2024.
- [60] F. Long, Z. Qiu, T. Yao, and T. Mei, “Videodrafter: Content-consistent multi-scene video generation with llm,” *arXiv preprint arXiv:2401.01256*, 2024.
- [61] Z. Wu, A. Siarohin, W. Menapace, I. Skorokhodov, Y. Fang, V. Chordia, I. Gilitschenski, and S. Tulyakov, “Mind the time: Temporally-controlled multi-event video generation,” *arXiv preprint arXiv:2412.05263*, 2024.
- [62] Z. Wang, J. Li, H. Lin, J. Yoon, and M. Bansal, “Dreamrunner: Fine-grained storytelling video generation with retrieval-augmented motion adaptation,” *arXiv preprint arXiv:2411.16657*, 2024.
- [63] S. Zhuang, K. Li, X. Chen, Y. Wang, Z. Liu, Y. Qiao, and Y. Wang, “Vlogger: Make your dream a vlog,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8806–8817.
- [64] Y. Li, H. Shi, B. Hu, L. Wang, J. Zhu, J. Xu, Z. Zhao, and M. Zhang, “Anim-director: A large multimodal model powered agent for controllable animation video generation,” in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.
- [65] T. Zhang, X. Xie, X. Du, and H. Xie, “Sketch-guided scene image generation,” *arXiv preprint arXiv:2407.06469*, 2024.
- [66] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, “Videocomposer: Compositional video synthesis with motion controllability,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [67] J. Li, J. Yang, A. Hertzmann, J. Zhang, and T. Xu, “Layoutgan: Generating graphic layouts with wireframe discriminators,” *arXiv preprint arXiv:1901.06767*, 2019.
- [68] J. Wu, C. Tang, J. Wang, Y. Zeng, X. Li, and Y. Tong, “Diffsensei: Bridging multi-modal llms and diffusion models for customized manga generation,” *arXiv preprint arXiv:2412.07589*, 2024.
- [69] S. Chen, D. Li, Z. Bao, Y. Zhou, L. Tan, Y. Zhong, and Z. Zhao, “Manga generation via layout-controllable diffusion,” *arXiv preprint arXiv:2412.19303*, 2024.
- [70] H. He, Y. Xu, Y. Guo, G. Wetzstein, B. Dai, H. Li, and C. Yang, “Cameractrl: Enabling camera control for text-to-video generation,” *arXiv preprint arXiv:2404.02101*, 2024.
- [71] L. Hu, X. Gao, P. Zhang, K. Sun, B. Zhang, and L. Bo, “Animate anyone: Consistent and controllable image-to-video synthesis for character animation,” *arXiv preprint arXiv:2311.17117*, 2023.
- [72] S. Zhu, J. L. Chen, Z. Dai, Y. Xu, X. Cao, Y. Yao, H. Zhu, and S. Zhu, “Champ: Controllable and consistent human image animation with 3d parametric guidance,” in *European Conference on Computer Vision (ECCV)*, 2024.
- [73] Y. Zhang, J. Gu, L.-W. Wang, H. Wang, J. Cheng, Y. Zhu, and F. Zou, “Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance,” *arXiv preprint arXiv:2406.19680*, 2024.
- [74] S. Tan, B. Gong, X. Wang, S. Zhang, D. Zheng, R. Zheng, K. Zheng, J. Chen, and M. Yang, “Animate-x: Universal character image animation with enhanced motion representation,” *arXiv preprint arXiv:2410.10306*, 2024.
- [75] J. Zhang, X. Zeng, X. Chen, W. Zuo, G. Yu, and Z. Tu, “Mikudance: Animating character art with mixed motion dynamics,” *arXiv preprint arXiv:2411.08656*, 2024.
- [76] Z. Lin, A. Huang, and Z. Huang, “Collaborative neural rendering using anime character sheets,” *arXiv preprint arXiv:2207.05378*, 2022.
- [77] H. Mo, C. Gao, and R. Wang, “Joint stroke tracing and correspondence for 2d animation,” *ACM Trans. Graph.*, vol. 43, no. 3, Apr. 2024.
- [78] L. Siyao, S. Zhao, W. Yu, W. Sun, D. Metaxas, C. C. Loy, and Z. Liu, “Deep animation video interpolation in the wild,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6587–6595.
- [79] T. Fukusato, A. Maejima, T. Igarashi, and T. Yotsukura, “Exploring inbetween charts with trajectory-guided sliders for cutout animation,” *Multimedia Tools and Applications*, vol. 83, no. 15, pp. 44 581–44 594, 2024.
- [80] W. Shen, C. Ming, W. Bao, G. Zhai, L. Chenn, and Z. Gao, “Enhanced deep animation video interpolation,” in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 31–35.
- [81] J. Shen, K. Hu, W. Bao, C. W. Chen, and Z. Wang, “Bridging the gap: Sketch-aware interpolation network for high-quality animation sketch inbetweening,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 10 287–10 295.
- [82] M. Tanveer, Y. Zhou, S. Niklaus, A. M. Amiri, H. Zhang, K. K. Singh, and N. Zhao, “Motionbridge: Dynamic video inbetweening with flexible controls,” *arXiv preprint arXiv:2412.13190*, 2024.
- [83] toona.io. [Online]. Available: <https://toona.io/>
- [84] X. Li, B. Zhang, J. Liao, and P. V. Sander, “Deep sketch-guided cartoon video inbetweening,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 8, pp. 2938–2952, 2021.
- [85] L. Siyao, T. Gu, W. Xiao, H. Ding, Z. Liu, and C. C. Loy, “Deep geometrized cartoon line inbetweening,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7291–7300.
- [86] Y. Meng, H. Ouyang, H. Wang, Q. Wang, W. Wang, K. L. Cheng, Z. Liu, Y. Shen, and H. Qu, “Anidoc: Animation creation made easier,” *arXiv preprint arXiv:2412.14173*, 2024.
- [87] S. Yang, L. Jiang, Z. Liu, and C. C. Loy, “Vtoonify: Controllable high-resolution portrait video style transfer,” *ACM Transactions on Graphics (TOG)*, vol. 41, no. 6, pp. 1–15, 2022.
- [88] S. Yang, L. Jiang, Z. Liu, , and C. C. Loy, “Styleganex: Stylegan-based manipulation beyond cropped aligned faces,” in *ICCV*, 2023.
- [89] S. Yang, Y. Zhou, Z. Liu, and C. C. Loy, “Fresco: Spatial-temporal correspondence for zero-shot video translation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8703–8712.
- [90] M. Geyer, O. Bar-Tal, S. Bagon, and T. Dekel, “Tokenflow: Consistent diffusion features for consistent video editing,” *arXiv preprint arXiv:2307.10373*, 2023.
- [91] M. Shi, J.-Q. Zhang, S.-Y. Chen, L. Gao, Y.-K. Lai, and F.-L. Zhang, “Deep line art video colorization with a few references,” *arXiv preprint arXiv:2003.10685*, 2020.
- [92] D. Loftsdóttir and M. Guzdić, “Sketchbetween: Video-to-video synthesis for sprite animation via sketches,” in *Proceedings of the 17th International Conference on the Foundations of Digital Games*, 2022, pp. 1–7.
- [93] N. Wang, M. Niu, Z. Dou, Z. Wang, Z. Wang, Z. Ming, B. Liu, and H. Li, “Coloring anime line art videos with transformation region enhancement network,” *Pattern Recognition*, vol. 141, p. 109562, 2023.
- [94] Y. Yu, J. Qian, C. Wang, Y. Dong, and B. Liu, “Animation line art colorization based on optical flow method,” 2022. [Online]. Available: <http://dx.doi.org/10.2139/ssrn.4202289>
- [95] Y. Yu, Z. Zeng, H. Hua, J. Fu, and J. Luo, “Promptfix: You prompt and we fix the photo,” *arXiv preprint arXiv:2405.16785*, 2024.
- [96] Z. Huang, M. Zhang, and J. Liao, “Lved: reference-based lineart video colorization with diffusion models,” *ACM Transactions on Graphics (TOG)*, vol. 43, no. 6, pp. 1–11, 2024.
- [97] E. Casey, V. Pérez, and Z. Li, “The animation transformer: Visual correspondence via segment matching,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 323–11 332.
- [98] Y. Dai, S. Zhou, Q. Li, C. Li, and C. C. Loy, “Learning inclusion matching for animation paint bucket colorization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 544–25 553.
- [99] Anonymous, “Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport,” in *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=u1cQYxR1H>
- [100] W. Cong, J. Zhang, L. Niu, L. Liu, Z. Ling, W. Li, and L. Zhang, “Dovenet: Deep image harmonization via domain verification,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8394–8403.

- [101] W. Cong, X. Tao, L. Niu, J. Liang, X. Gao, Q. Sun, and L. Zhang, “High-resolution image harmonization via collaborative dual transformations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 470–18 479.
- [102] J. J. A. Guerreiro, M. Nakazawa, and B. Stenger, “Pct-net: Full resolution image harmonization using pixel-wise color transformations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5917–5926.
- [103] Y. Jiang, H. Zhang, J. Zhang, Y. Wang, Z. Lin, K. Sunkavalli, S. Chen, S. Amirghodsi, S. Kong, and Z. Wang, “Ssh: A self-supervised framework for image harmonization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4832–4841.
- [104] G. C. Tarrés, Z. Lin, Z. Zhang, J. Zhang, Y. Song, D. Ruta, A. Gilbert, J. Collomosse, and S. Y. Kim, “Thinking outside the bbox: Unconstrained generative object compositing,” *arXiv preprint arXiv:2409.04559*, 2024.
- [105] Y. Sheng, Z. Yu, L. Ling, Z. Cao, X. Zhang, X. Lu, K. Xian, H. Lin, and B. Benes, “Dr. bokeh: Differentiable occlusion-aware bokeh rendering,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4515–4525.
- [106] Y. Man, Y. Sheng, J. Zhang, L.-Y. Gui, and Y.-X. Wang, “Floating no more: Object-ground reconstruction from a single image,” *arXiv preprint arXiv:2407.18914*, 2024.
- [107] D. Winter, M. Cohen, S. Fruchter, Y. Pritch, A. Rav-Acha, and Y. Hoshen, “Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion,” *arXiv preprint arXiv:2403.18818*, 2024.
- [108] P. Sharma, V. Jampani, Y. Li, X. Jia, D. Lagun, F. Durand, B. Freeman, and M. Matthews, “Alchemist: Parametric control of material properties with diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 130–24 141.
- [109] H. Chen, X. Wang, Y. Zhang, Y. Zhou, Z. Zhang, S. Tang, and W. Zhu, “Disenstudio: Customized multi-subject text-to-video generation with disentangled spatial control,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 3637–3646.
- [110] Y. Sheng, J. Zhang, and B. Benes, “Ssn: Soft shadow network for image compositing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4380–4390.
- [111] Y. Tang, S. Xu, T. Wang, Q. Lin, Q. Lu, and F. Zheng, “Multi-modal segment assemblage network for ad video editing with importance-coherence reward,” in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2022, pp. 3519–3535.
- [112] J. Cao, Y. Wu, W. Chi, W. Zhu, Z. Su, and J. Wu, “Reframe anything: Llm agent for open world video reframing,” *arXiv preprint arXiv:2403.06070*, 2024.
- [113] OpusClip. [Online]. Available: <https://www.opus.pro>
- [114] C. Gan, D. Huang, P. Chen, J. B. Tenenbaum, and A. Torralba, “Foley music: Learning to generate music from videos,” in *ECCV*, 2020.
- [115] J. Kang, S. Poria, and D. Herremans, “Video2music: Suitable music generation from videos using an affective multimodal transformer model,” *Expert Systems with Applications*, p. 123640, 2024.
- [116] K. Su, J. Y. Li, Q. Huang, D. Kuzmin, J. Lee, C. Donahue, F. Sha, A. Jansen, Y. Wang, M. Verzetti *et al.*, “V2meow: Meowing to the visual beat via video-to-music generation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4952–4960.
- [117] S. Chowdhury, S. Nag, J. K. J. B. Vasan Srinivasan, and D. Manocha, “Melfusion: Synthesizing music from image and language cues using diffusion models,” *CVPR*, 2024.
- [118] Z. Tian, Z. Liu, R. Yuan, J. Pan, Q. Liu, X. Tan, Q. Chen, W. Xue, and Y. Guo, “Vidmuse: A simple video-to-music generation framework with long-short-term modeling,” *arXiv preprint arXiv:2406.04321*, 2024.
- [119] V. Iashin and E. Rahtu, “Taming visually guided sound generation,” in *British Machine Vision Conference (BMVC)*, 2021.
- [120] R. Sheffer and Y. Adi, “I hear your true colors: Image guided audio generation,” 2022.
- [121] X. Mei, V. Nagaraja, G. Le Lan, Z. Ni, E. Chang, Y. Shi, and V. Chandra, “Foleygen: Visually-guided audio generation,” in *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2024, pp. 1–6.
- [122] S. Luo, C. Yan, C. Hu, and H. Zhao, “Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [123] C. Chen, P. Peng, A. Baid, Z. Xue, W.-N. Hsu, D. Harwath, and K. Grauman, “Action2sound: Ambient-aware generation of action sounds from egocentric videos,” *arXiv preprint arXiv:2406.09272*, 2024.
- [124] G. Cong, Y. Qi, L. Li, A. Beheshti, Z. Zhang, A. v. d. Hengel, M.-H. Yang, C. Yan, and Q. Huang, “Styledubber: Towards multi-scale style learning for movie dubbing,” *arXiv preprint arXiv:2402.12636*, 2024.
- [125] K. Cai, C. Liu, and D. M. Chan, “Anim-400k: A large-scale dataset for automated end to end dubbing of video,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1–5.
- [126] G. Cong, J. Pan, L. Li, Y. Qi, Y. Peng, A. van den Hengel, J. Yang, and Q. Huang, “Emodubber: Towards high quality and emotion controllable movie dubbing,” 2024.
- [127] Z. Zhang, L. Li, G. Cong, H. YIN, Y. Gao, C. Yan, A. van den Hengel, and Y. Qi, “From speaker to dubber: Movie dubbing with prosody and duration consistency learning,” in *ACM Multimedia 2024*, 2024.
- [128] G. Cong, L. Li, Y. Qi, Z.-J. Zha, Q. Wu, W. Wang, B. Jiang, M.-H. Yang, and Q. Huang, “Learning to dub movies via hierarchical prosody models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 687–14 697.
- [129] Q. Chen, M. Tan, Y. Qi, J. Zhou, Y. Li, and Q. Wu, “V2c: Visual voice cloning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 242–21 251.
- [130] J. L. Cardoso, F. Banterle, P. Cignoni, and M. Wimmer, “Re: Draw-context aware translation as a controllable method for artistic production,” *arXiv preprint arXiv:2401.03499*, 2024.
- [131] C. Huang, S. Liang, Y. Tang, Y. T. A. Kumar, and C. Xu, “Scaling concept with text-guided diffusion models,” 2024.
- [132] L. Zhang, T.-T. Wong, and Y. Liu, “Sprite-from-sprite: Cartoon animation decomposition with self-supervised sprite estimation,” *ACM Trans. Graph.*, vol. 41, no. 6, Nov. 2022.
- [133] Y.-C. Lee, E. Lu, S. Rumbley, M. Geyer, J.-B. Huang, T. Dekel, and F. Cole, “Generative omnimatte: Learning to decompose video into layers,” *arXiv preprint arXiv:2411.16683*, 2024.
- [134] E. Weber, R. Peterlinz, R. Mathur, F. Warburg, A. A. Efros, and A. Kanazawa, “Toon3d: Seeing cartoons from a new perspective,” *arXiv preprint arXiv:2405.10320*, 2024.
- [135] J. Zhou, C. Xiao, M.-L. Lam, and H. Fu, “Drawingspinup: 3d animation from single character drawings,” in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–10.
- [136] F. Shen, H. Ye, S. Liu, J. Zhang, C. Wang, X. Han, and W. Yang, “Boosting consistency in story visualization with rich-contextual conditional diffusion models,” *arXiv preprint arXiv:2407.02482*, 2024.
- [137] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman, “Dreampose: Fashion image-to-video synthesis via stable diffusion,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 22 623–22 633.
- [138] Z. Tong, C. Li, Z. Chen, B. Wu, and W. Zhou, “Musepose: a pose-driven image-to-video framework for virtual human generation,” *arxiv*, 2024.
- [139] L. Song, P. Liu, G. Yin, and C. Xu, “Adaptive super resolution for one-shot talking-head generation,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4115–4119.
- [140] L. Song, B. Liu, and N. Yu, “Talking face video generation with editable expression,” in *Image and Graphics: 11th International Conference, ICIG 2021, Haikou, China, August 6–8, 2021, Proceedings, Part III 11*. Springer, 2021, pp. 753–764.
- [141] P. Liu, L. Song, D. Zhang, H. Hua, Y. Tang, H. Tu, J. Luo, and C. Xu, “Emo-avatar: Efficient monocular video style avatar through texture rendering,” *arXiv preprint arXiv:2402.00827*, 2024.
- [142] S. Yang, Y. Zhou, Z. Liu, , and C. C. Loy, “Rerender a video: Zero-shot text-guided video-to-video translation,” in *ACM SIGGRAPH Asia Conference Proceedings*, 2023.
- [143] Y. Song, Z. Zhang, Z. Lin, S. Cohen, B. Price, J. Zhang, S. Y. Kim, and D. Aliaga, “Objectstitch: Object compositing with diffusion model,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 18 310–18 319.
- [144] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumley, “AudioLDM: Text-to-audio generation with latent diffusion models,” *Proceedings of the International Conference on Machine Learning*, pp. 21 450–21 474, 2023.
- [145] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumley, “Audiodlm 2: Learning holistic audio generation with self-supervised pretraining,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.

- [146] Q. Huang, D. S. Park, T. Wang, T. I. Denk, A. Ly, N. Chen, Z. Zhang, Z. Zhang, J. Yu, C. Frank *et al.*, “Noise2music: Text-conditioned music generation with diffusion models,” *arXiv preprint arXiv:2302.03917*, 2023.
- [147] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, “Musiclm: Generating music from text,” *arXiv preprint arXiv:2301.11325*, 2023.
- [148] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, “Simple and controllable music generation,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [149] R. Huang, M. Li, D. Yang, J. Shi, X. Chang, Z. Ye, Y. Wu, Z. Hong, J. Huang, J. Liu *et al.*, “Audiogpt: Understanding and generating speech, music, sound, and talking head,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, 2024, pp. 23 802–23 804.
- [150] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu, “Language-guided joint audio-visual editing via one-shot adaptation,” in *Proceedings of the Asian Conference on Computer Vision*, 2024, pp. 1011–1027.
- [151] Z. Pan, Y. Zhu, and Y. Mu, “Sakuga-42m dataset: Scaling up cartoon research,” *arXiv preprint arXiv:2405.07425*, 2024.
- [152] Z. Guo, H. Zheng, Y. Jiang, Z. Gu, and B. Zheng, “Intrinsic image harmonization,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 367–16 376.
- [153] L. Siyao, Y. Li, B. Li, C. Dong, Z. Liu, and C. C. Loy, “Animerun: 2d animation visual correspondence from open source 3d movies,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 996–19 007, 2022.
- [154] AnitaTeam, “Anita dataset: An industrial animation dataset.” [Online]. Available: https://zhenglinpan.github.io/AnitaDataset_homepage/
- [155] H. Hua, Q. Liu, L. Zhang, J. Shi, Z. Zhang, Y. Wang, J. Zhang, and J. Luo, “Finecaption: Compositional image captioning focusing on wherever you want at any granularity,” *arXiv preprint arXiv:2411.15411*, 2024.
- [156] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang, “Mm-vet: Evaluating large multimodal models for integrated capabilities,” *arXiv preprint arXiv:2308.02490*, 2023.
- [157] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun *et al.*, “Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9556–9567.
- [158] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin *et al.*, “Are we on the right way for evaluating large vision-language models?” *arXiv preprint arXiv:2403.20330*, 2024.
- [159] H. Hua, Y. Tang, Z. Zeng, L. Cao, Z. Yang, H. He, C. Xu, and J. Luo, “Mmcomposition: Revisiting the compositionality of pre-trained vision-language models,” *arXiv preprint arXiv:2410.09733*, 2024.
- [160] Y. Tang, J. Guo, H. Hua, S. Liang, M. Feng, X. Li, R. Mao, C. Huang, J. Bi, Z. Zhang *et al.*, “Vidcomposition: Can milms analyze compositions in compiled videos?” *arXiv preprint arXiv:2411.10979*, 2024.
- [161] C. Zhou, L. Yu, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, J. Kahn, X. Ma, L. Zettlemoyer, and O. Levy, “Transfusion: Predict the next token and diffuse images with one multi-modal model,” *arXiv preprint arXiv:2408.11039*, 2024.
- [162] L. Yu, B. Shi, R. Pasunuru, B. Muller, O. Golovneva, T. Wang, A. Babu, B. Tang, B. Karrer, S. Sheynin *et al.*, “Scaling autoregressive multi-modal models: Pretraining and instruction tuning,” *arXiv preprint arXiv:2309.02591*, vol. 2, no. 3, 2023.
- [163] E. Aiello, L. Yu, Y. Nie, A. Aghajanyan, and B. Oguz, “Jointly training large autoregressive multimodal models,” *arXiv preprint arXiv:2309.15564*, 2023.
- [164] C. Team, “Chameleon: Mixed-modal early-fusion foundation models, 2024,” URL <https://arxiv.org/abs/2405.09818>.
- [165] J. An, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, L. Wang, and J. Luo, “Openleaf: Open-domain interleaved image-text generation and evaluation,” *arXiv preprint arXiv:2310.07749*, 2023.
- [166] C. Tian, X. Zhu, Y. Xiong, W. Wang, Z. Chen, W. Wang, Y. Chen, L. Lu, T. Lu, J. Zhou *et al.*, “Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer,” *arXiv preprint arXiv:2401.10208*, 2024.
- [167] Z. Tang, Z. Yang, M. Khademi, Y. Liu, C. Zhu, and M. Bansal, “Codi-2: In-context interleaved and interactive any-to-any generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 425–27 434.
- [168] T. Xie, F. Zhou, Z. Cheng, P. Shi, L. Weng, Y. Liu, T. J. Hua, J. Zhao, Q. Liu, C. Liu *et al.*, “Openagents: An open platform for language agents in the wild,” *arXiv preprint arXiv:2310.10634*, 2023.
- [169] H. Hua, J. Shi, K. Kafle, S. Jenni, D. Zhang, J. Collomosse, S. Cohen, and J. Luo, “Finematch: Aspect-based fine-grained image and text mismatch detection and correction,” in *European Conference on Computer Vision*. Springer, 2025, pp. 474–491.
- [170] S. Lin, W. Hua, L. Li, C.-J. Chang, L. Fan, J. Ji, H. Hua, M. Jin, J. Luo, and Y. Zhang, “Battleagent: Multi-modal dynamic emulation on historical battles to complement historical analysis,” *arXiv preprint arXiv:2404.15532*, 2024.
- [171] S. Hong, X. Zheng, J. Chen, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou *et al.*, “Metagpt: Meta programming for multi-agent collaborative framework,” *arXiv preprint arXiv:2308.00352*, 2023.
- [172] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian *et al.*, “Toolllm: Facilitating large language models to master 16000+ real-world apis,” *arXiv preprint arXiv:2307.16789*, 2023.
- [173] T. Wang, J. Zhang, J. Fei, H. Zheng, Y. Tang, Z. Li, M. Gao, and S. Zhao, “Caption anything: Interactive image description with diverse multimodal controls,” *arXiv preprint arXiv:2305.02677*, 2023.
- [174] M. Li, Y. Zhao, B. Yu, F. Song, H. Li, H. Yu, Z. Li, F. Huang, and Y. Li, “Api-bank: A comprehensive benchmark for tool-augmented llms,” *arXiv preprint arXiv:2304.08244*, 2023.
- [175] A. Gu, B. Rozière, H. Leather, A. Solar-Lezama, G. Synnaeve, and S. I. Wang, “Cruxeval: A benchmark for code reasoning, understanding and execution,” *arXiv preprint arXiv:2401.03065*, 2024.
- [176] X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang *et al.*, “Agentbench: Evaluating llms as agents,” *arXiv preprint arXiv:2308.03688*, 2023.

APPENDIX

A. Cel-Animations Featured in the Teaser

From left to right, the works are:

- *Snow White and the Seven Dwarfs* (1937)
- *Fantasia* (1940)
- *Castle in the Sky* (1986)
- *Beauty and the Beast* (1991)
- *Neon Genesis Evangelion* (1995)
- *Shirobako* (2014)
- *Demon Slayer: Kimetsu no Yaiba* (2019)
- *Oshi no Ko* (2023)
- *The Dog & The Boy* (2022)
- *Rock, Paper, Scissors* (2023)
- *Twins Hinahima* (scheduled for release in Spring 2025)

B. Details about Quality Control & Inspection

Several inspection stages ensure animation quality throughout production:

- *Animation Check*: Animation directors review and correct keyframes to maintain character consistency and motion quality.
- *Time Sheet Check*: Timing directors verify movement rhythm and frame timing.
- *Trace & Paint Check*: Verification of line quality and color accuracy before final composition.
- *Final Check*: Overall quality inspection of completed cuts, including photography and effects.

C. Terminology

Term	Description ⁵	Illustration
3D Assistance	Place buildings and characters in 3D space to recreate the actual scene, and position a virtual camera within the 3D space to finalize the layout. The layout created in 3D is printed onto paper, and from there, the process of adding character movements by hand continues.	
AR	The abbreviation for “after recording.” It refers to the process of recording audio to match the visuals. In the case of animation, it often uses footage specifically prepared for after-recording rather than the final version of the visuals. Conversely, the method of creating visuals to match pre-recorded audio is called “prescoring” (pre-scoring).	
Background (BG)	Refers to background materials or the department responsible for them. These are primarily illustrations placed beneath cels to depict the location, setting, emotions, speed, and other elements. They are traditionally painted on drawing paper using poster colors. With the advent of digitalization, backgrounds painted on drawing paper began to be scanned, and in recent years, more studios have started creating them directly as digital data on computers.	
Book	The term refers to a background layer that is not placed at the very bottom during filming, despite being drawn separately on another sheet for a different video. This layer is stacked on top of the cel but is created using the same drawing technique as a regular background (BG). ⁸	
Color Design	Color design creates “color models” for each scene, which detail the colors used to paint characters.	
Color Direction	The color direction process involves determining which color model to use for each cut and directing its application. For small items that appear only in a single episode, the color director often decides their colors. Once the direction is complete, the cuts are sent to the “finishing” section, where the actual coloring is done. During inspection, it is checked whether the directed colors have been applied accurately, and any coloring mistakes or omissions are identified.	
Compositing	The process of animating static parts on a separate sheet and then compositing them with the moving parts during the finalization stage to create a single image.	

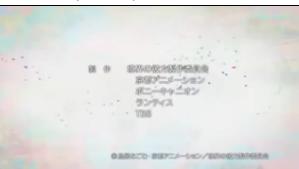
⁵<https://msc-jp.biz/material/words/>; <http://shirobako-anime.com/words.html>

⁶<https://cgworld.jp/feature/202111-atsuc2021-shaft.html>

⁷<https://www.clipstudio.net/en/animation/skills-knowledge/>

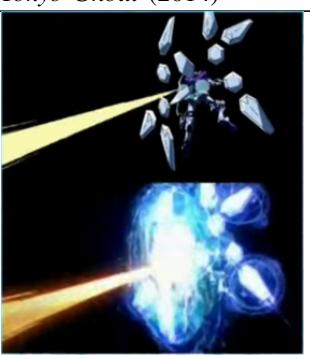
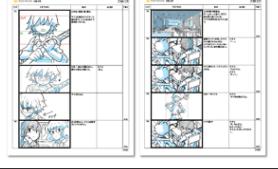
⁸<https://w.atwiki.jp/aniken/>

⁹https://x.com/st_bind/status/1598965269459779585

Cutting (CT)	Refers to re-editing the developed rush film to match the storyboard sequence. At this stage, the cut lengths are adjusted by deciding on the timing (duration in seconds) in consultation with the director. For film works, this process involves physically cutting and pasting the rush film, hence the name. However, the term “CT” (cutting) is often used to refer to editing performed before after-recording (AR). Nowadays, as most works are digital, editing is primarily done digitally on computers.	
Dubbing (DB)	The process of synchronizing audio data, background music (BGM), and sound effects with the visuals.	
Genga	Based on the layout, it depicts the foundational movements that serve as the basis for the animation. These are drawn by key animators.	 <p>Boruto: Naruto Next Generations (2018)</p>
Ending (ED)	The section that signifies the conclusion of a work. It typically features theme music and visuals while displaying the names of the staff and cast involved in the production.	 <p>Beyond the Boundary (2013)</p>
Inbetweening	One of the fundamental tasks in animation, involving drawing the in-between lines of movement between two frames. It can also refer to the in-between frames that connect one keyframe to another.	 <p>Keyframe Keyframe Keyframe</p> <p>10</p>
Layout (L/O)	The composition of the screen is based on the storyboard, depicting the arrangement of characters and the background. Once the layout direction check and layout animation director check are completed, this layout is handed over to the background section to request background artwork production.	 <p>Your Lie in April (2014)</p>
Mood Board	A mood-board's purpose is to capture the ambiance of the project with a collection of images, thoughts, color panels, and general design ideas. Many designers use mood boards in different ways, but they all have the same goal, create a visualization of the concept before production to convey the feel of the project. ¹¹	 <p>11</p>
Nigen	An abbreviation for “Dainigenga”, also called “second key animation”. The key animation process is divided into two stages: layout and key animation. Sometimes, the key animation stage is further divided into first key animation and second key animation. Second key animators serve as assistants to the first key animators.	 <p>Iron Saga (2019)</p>

¹⁰<https://www.pixivision.net/en/a/800>

¹¹[https://www.yansmedia.com/blog/what-is-moodboard?](https://www.yansmedia.com/blog/what-is-moodboard/)

Opening (OP)	The section that marks the beginning of a work. It typically features visuals that convey the essence of the work, accompanied by the title, theme music, and credits introducing the main staff.	 Tokyo Ghoul (2014)
Photography	In the modern era of digitalization, this refers to the section that composites background and character assets using software, applies various camera effects, and converts them into movie data. Before the transition to digital, cels were layered over backgrounds painted on paper and physically filmed with a camera.	
Script	A description of when, where, who, and what is happening. It documents the characters' actions and dialogues in text form, consisting of scene headings, dialogue, and stage directions.	
SE	An abbreviation for “Sound Effects”. It refers to sound effects other than dialogue and music.	
Storyboard	Based on the script, this document includes cut divisions, screen compositions, character actions, dialogue, and timing, all annotated with illustrations. It can be described as the blueprint for an animation’s visuals.	 12
VTR	An abbreviation for “Video Tape Recorder”.	

¹²<https://blog.toonboom.com/ja>