Create AI

WHITE PAPER ON THE APPLICATION AND DEVELOPMENT OF MULTIMODAL GENERATIVE TECHNOLOGY IN ANIMATION PRODUCTION

May, 2025

INTRODUCTION

At a time when digitalization and artificial intelligence are reshaping industry fundamentals, the animation production industry is undergoing profound changes centered around Multimodal Generative Technology. Emblematic of cutting-edge technology with its cross-modal data processing capability, this technology achieves deep analysis and cross-modal reconstruction of unstructured data, such as text, images, audio, and video through data alignment and joint modeling mechanisms. This technological breakthrough not only significantly enhances content production efficiency but also expands the boundaries of artistic expression through multimodal collaborative creation, injecting strong momentum into the full-process innovation of the animation industry. Notably, according to IDC, by 2028 unstructured data will account for 82.3% of all data, and Multimodal Generative Technology, with its unique advantages in handling complex data types, is quickly becoming the core technological foundation driving the industry's leap in efficiency and quality.

The practical value of this technological transformation has recently been fully validated by the groundbreaking 2025 animated film "Nezha: The Demon Child Conquers the Sea" (hereinafter referred to as "Nezha 2"). The production team behind "Nezha 2" reconstructed 60% of the industrial process by building a Multimodal Generative Technology stack, achieving a 30% reduction in production cycle; 470% increase in per capita productivity; 20% reduction in costs; and, a 40% improvement in picture quality. The results are reflected in the global box office exceeding RMB 13.9 billion, with a peak daily box office of RMB 700 million (as of February 28, 2025), setting a new industry record. The film's technological applications cover the entire chain from art design and motion capture to multilingual promotion, becoming a milestone case in the AI-driven industrialization of art.

Based on these technological evolution and practical breakthroughs, this White Paper will delve into the application scenarios, core technologies, challenges, and future development paths of Multimodal Generative Technology in animation production, combined with authoritative third-party data to dissect benchmark cases such as "Nezha 2," ultimately serving as a valuable and comprehensive reference for industry practitioners, researchers, and other stakeholders.

CONTENTS

INTRODUCTION

FULL PROCESS OF ANIMATION PRODUCTION AND APPLICATION SCENARIOS OF MULTIMODAL GENERATIVE TECHNOLOGY

01

1 | OVERVIEW OF THE FULL PROCESS OF ANIMATION PRODUCTION
2 | APPLICATION OF MULTIMODAL GENERATIVE TECHNOLOGY IN THE CORE PRODUCTION STAGE

ESTABLISHING EVALUATION STANDARDS FOR VIDEO MODELS BASED ON ANIMATION APPLICATION SCENARIOS

02

INDUSTRIAL-GRADE IMAGE-TO-VIDEO MODEL "RUYI" - CUSTOM DEVELOPMENT TO BALANCE SCENARIO NEEDS

03

1 | BACKGROUND OF CUSTOM DEVELOPMENT
2 | DATA-LEVEL OPTIMIZATION
3 | INDUSTRIAL-GRADE USAGE STANDARDS

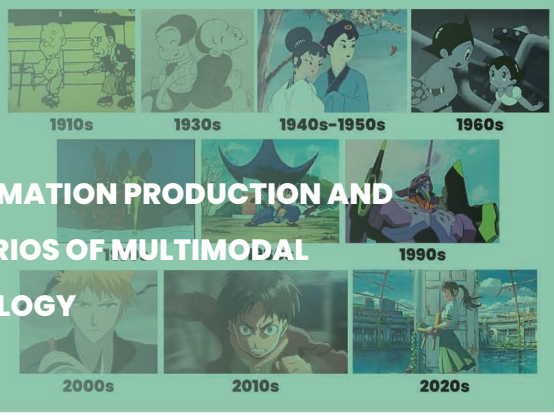
CHALLENGES AND PROSPECTS OF MULTIMODAL GENERATIVE TECHNOLOGY IN ANIMATION PRODUCTION

04

1 | CHALLENGES
2 | FUTURE PROSPECTS

CONCLUSION

APPENDIX : DATA SOURCE DESCRIPTION



01 FULL PROCESS OF ANIMATION PRODUCTION AND APPLICATION SCENARIOS OF MULTIMODAL GENERATIVE TECHNOLOGY

OVERVIEW OF THE FULL PROCESS OF ANIMATION PRODUCTION

"From an idea' s conception to the final presentation to an audience, animation production is a complex and intricate process that includes five closely connected stages: project initiation, pre-production, production, post-production, and distribution and marketing. Each stage plays an indispensable role in shaping the quality, market impact, and audience feedback of the final animation work."

			Core Challenges	Training Data	Core Architecture
Project Initiation	Script Finalization Team Selection Obtaining Permits Securing Investment				
Pre-production	Concept Finalization Storyboarding Animatics layout	Large Language Model (LLM)	1. Contextual Understanding: Capturing long-distance dependencies and understanding textual meaning 2. Generalization Ability: Applicable to various language tasks, including reasoning, summarization, and dialogue 3. Token Prediction: Learning and generating the next most likely token	Text Corpus	Transformer
		Text-to-Image Model (Text-to-Image)	1. Visual Consistency: Ensuring generated images match the input text 2. Style and Quality: Balancing artistic effects and realism	Image-Text Paired Dataset	Diffusion Transformer,GAN,Cross-modal Transformer
Production	Asset Creation Animation Lighting and Effects Compositing	Image-to-Video Model (I2V) (Frame Sequence)	1. Consistency: Ensuring smooth and consistent motion between frames while maintaining object shape and style 2. Motion Prediction: Predicting reasonable motion patterns based on input images, such as flowing water, walking people, etc 3. Frame Rate and Resolution: Ensuring stable frame rates while generating high-resolution videos	Video Dataset	CNN + RNN/ Transformer,or Diffusion Transformer
Post-production	Voiceover Color Grading Subtitles	Audio Generation Model	1. Naturalness, fluency, and intelligibility of speech synthesis 2. Matching sound effects with scenes	Sound Dataset	CNN，GAN
Marketing & Distribution	Obtaining Review Permits Marketing and Distribution Determining Platform Channels				

The project initiation stage is the starting point of any animation project, where key elements such as commercial goals, audience positioning, core creativity, and budget planning need to be clarified, laying a solid foundation for subsequent production stages. The pre-production, production, and post-production stages together constitute the key stages of transforming creative ideas into specific audiovisual works, which we define as the core production stage. The distribution and marketing stage determines whether the work can be successfully launched into the market and recognized by the audience, expanding the work's influence and dissemination through effective promotional strategies and distribution channel planning.

In this White Paper, given the most significant application value and innovation potential of Multimodal Generative Technology in the core production stage, the subsequent sections will focus on this stage, providing a detailed explanation of how this technology deeply empowers the animation production process.

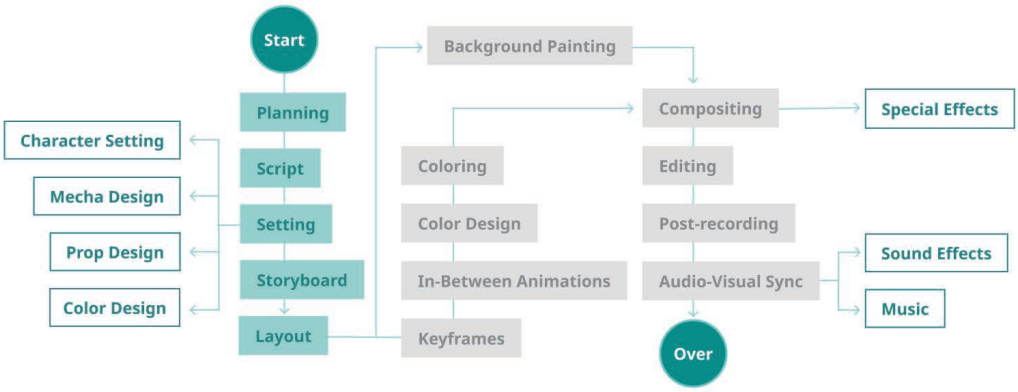
APPLICATION OF MULTIMODAL GENERATIVE TECHNOLOGY IN THE CORE PRODUCTION STAGE

In the current era of accelerated restructuring of digital content production paradigms, the application of Multimodal Generative Technology in 2D and 3D animation varies significantly. The application in 2D animation is more mature and widespread. However, from a medium to long-term perspective of three to five years, 3D animation presents greater disruptive opportunities. Below, we will explore the application of Multimodal Generative Technology in the core production stages of 2D and 3D animation.

2D Animation

■ Pre-production

The pre-production phase of a 2D animation project is primarily handled by the core creative team, including the animation director, screenwriter, art director (including character/scene designers), and storyboard artist. This phase covers planning, scripting, setting, storyboarding, and layout design.



Planning : This involves the project's initiation, including determining the title, style, type, number of episodes, and story content. Multimodal Generative Technology can analyze data from platforms like MyAnimeList and Bilibili, assessing genre popularity and user preferences, and match features of works by directors like Hayao Miyazaki and Makoto Shinkai to generate hybrid style proposals.

Scripting : Creating the story for each episode, typically involving one to three screenwriters. A "script coordinator" often oversees the story outline and each episode's script. Multimodal Generative Technology can expand plot branches using story tree algorithms, suggest episode pacing, monitor timeline inconsistencies, and recommend dialogue optimization.

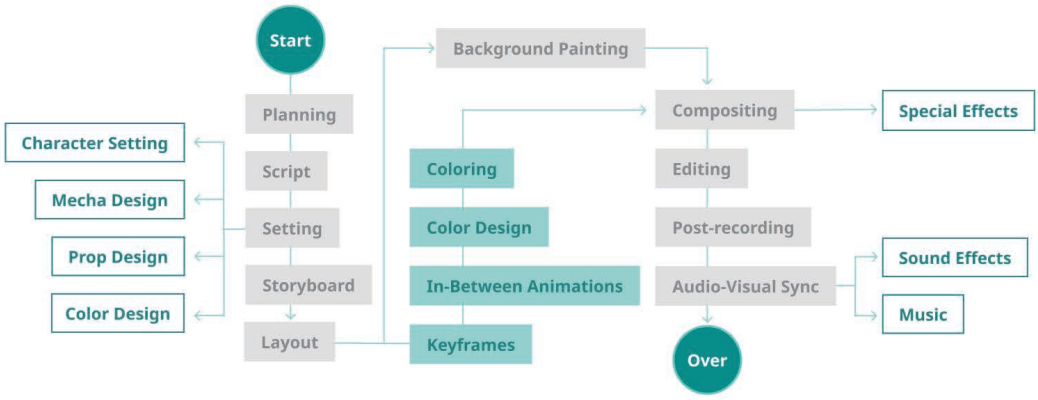
Setting : This includes concept design, character setting, mechanical setting, art setting, and color design. Open-source tools like Stable Diffusion, through LoRA model training, can achieve personalized style customization for 2D characters (e.g., cel-shading, watercolor texture). In terms of color design, Multimodal Generative Technology can generate color specification tables and match material textures.

Storyboarding : Translating the script into visual scenes. By inputting text descriptions of scene actions or character emotions, Multimodal Generative Technology can automatically generate storyboard sketches and camera movement plans, indicating duration and camera movement methods. If reference videos are uploaded, it can extract camera movement trajectories and action rhythms, converting them into storyboard language.

Layout Design : Based on storyboard sketches, this involves designing the exact positioning of characters, background details, character actions and expressions, action decomposition methods, and specific camera usage. Multimodal Generative Technology can generate character design line drafts through text descriptions and automatically match scene design elements that fit the world setting. By inputting reference materials, this technology can extract light and shadow styles, perspective rules, and other features, automatically generating design drafts that comply with physical laws, and provide dynamic adjustment suggestions for character and scene proportions.

■ Production

The production phase is mainly handled by the execution team, including key animators, in-between animators, color designers, and background artists. This phase covers key animation, in-between animation, color specification, and layered coloring.



Key Animation : Drawing key frames based on layout designs, transforming static images into motion. Multimodal Generative Technology can assist in generating key frames by learning character settings and motion rules, automatically producing key animation drafts that meet storyboard requirements.

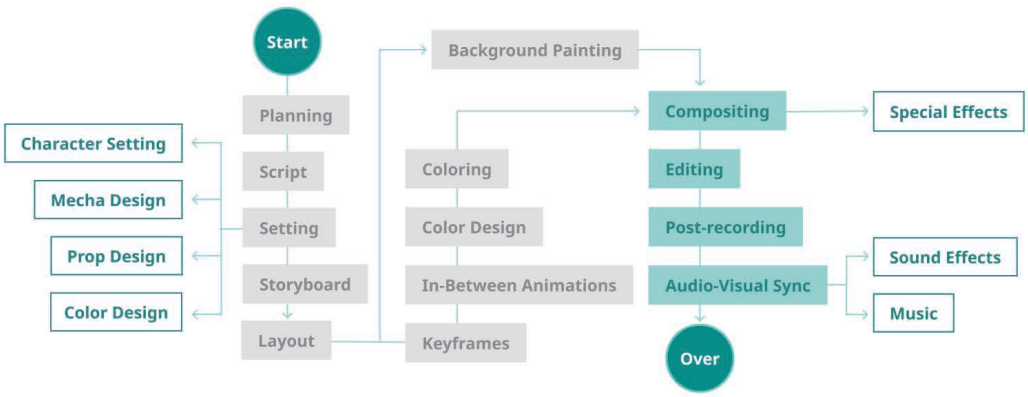
In-between Animation : Key animation defines key dynamic frames (Key Frames), while in-between animation (In-Between) fills in transitional frames to achieve smooth animation. Current mainstream Multimodal Generative Technology has achieved key frame interpolation and smooth transitional frame generation. However, there are still technical shortcomings in cross-frame color consistency control, and optimal continuation key frame selection based on temporal context remains a critical technical bottleneck.

Color Specification : Responsible for specifying the colors for each episode, particularly the dynamic parts, excluding the background. Multimodal Generative Technology can analyze scene atmosphere and character traits to generate color schemes that meet art director requirements, significantly reducing trial-and-error costs. This technology can also identify color contrast relationships between characters and backgrounds, automatically adjusting hue and saturation to enhance visual harmony.

Layered Coloring : Coloring the line drafts of key and in-between animations, excluding the background. Traditionally, after key animation is completed, the coloring phase strictly follows the color palette to ensure uniform colors for characters and props across different shots. Multimodal Generative Technology (e.g., segmentation networks) can accurately identify closed line draft areas, enabling one-click coloring and avoiding overflow issues in traditional processes. This technology also supports multiple coloring schemes for the same line draft, such as switching between day and night scenes, facilitating quick comparison and selection by the team.

■ Post-production

The post-production phase is mainly handled by the technical team, including compositors, editors, sound engineers, and voice directors. This phase includes animation compositing, shot editing, post-recording, and audio-visual synchronization.



Compositing : Combining key animation, in-between animation, background, CG, and other materials into one frame, adding special effects and camera movements as required by the storyboard. Multimodal Generative Technology can analyze frame content through algorithms, automatically generating smooth transition effects, such as matching transition methods based on scene tones or motion trajectories, reducing manual adjustment time. Based on deep learning, Multimodal Generative Technology can also repair low-resolution materials, optimize line smoothness, or automatically complete complex backgrounds, enhancing overall visual quality.

Editing : Connecting all shots into a complete animation, including adjusting shot order, adding or removing shot time, and eliminating unnecessary shots. Multimodal Generative Technology can analyze animation storyboard scripts and music rhythms, automatically recommending editing points to optimize narrative pacing. In terms of scene classification and tagging, Multimodal Generative Technology uses image recognition to classify and tag scenes and characters in the material library, quickly retrieving required clips and improving editing efficiency. This technology can also generate preliminary edited versions based on scripts or storyboards for further manual adjustments, shortening production cycles.

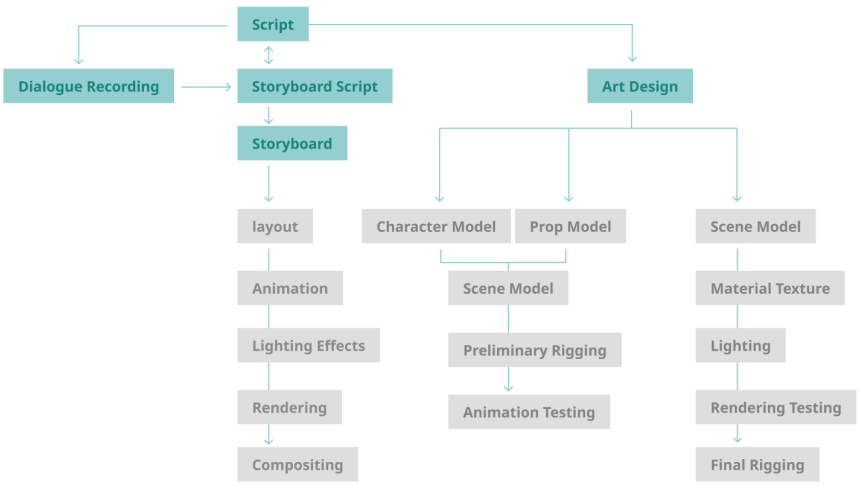
Post-recording : After the visuals are completed, voice actors perform dubbing. Multimodal Generative Technology supports generating natural voice dubbing and adjusting tone and emotion to match character traits, reducing reliance on professional voice actors. In terms of multilingual adaptation, Multimodal Generative Technology can quickly generate multilingual dubbing versions and adjust lip-sync animation to match speech through lip-sync driving technology, lowering localization costs.

Audio-visual Synchronization : Adding sound effects, BGM, and insert songs. Through timeline analysis, Multimodal Generative Technology can automatically align voice, sound effects, and visual actions, avoiding audio-visual synchronization issues. Based on scene emotions, this technology can dynamically adjust the mix ratio of background music, sound effects, and voice to enhance the atmosphere.

3D Animation

■ Pre-production

The 3D Animation pre-production of an animation project is mainly handled by the creative team, including the animation director, screenwriter, storyboard artist, character designer, art director, and storyboard artist. This covers four parts: creative finalization, art design, voice recording, and storyboard production.



Script Creation

Script creation is the source of the entire production. During traditional creation modes, screenwriters need long periods of time for brainstorming, data collection, and repeated polishing of plots and dialogues, easily falling into fixed thinking patterns.

With the introduction of Multimodal Generative Technology, creators provide original theme keywords and story outlines, and large models can supplement and generate logical scripts, greatly reducing repetitive labor for screenwriters and enriching creative inspiration.

The AI-assisted screenwriting system used in "Nezha 2" automatically generated plot branches that fit the mythological logic by analyzing millions of words of text from the Fengshen universe. The production team used ChatGPT to simulate "Wong Kar-wai-style dialogue" and "Nolan-style narrative structure," even generating Blender script codes to quickly test different artistic styles. The cyberpunk elements of the "Metaverse Heavenly Court" in the film were selected by AI from over 300 schemes.

Art Design

After finalizing the script, the art team needs to complete the design of characters, props, and scenes and organize them into a visual style guide for reference by other departments.

Previously, the art team needed to hand-draw a large number of concept images. Now, with the help of image generation models, they can quickly produce diverse and high-quality visual material references based on text descriptions. Creators can also achieve personalized customization through parameter adjustments to meet the needs of different animation projects.

In the creation of concept art for "**Nezha 2**",

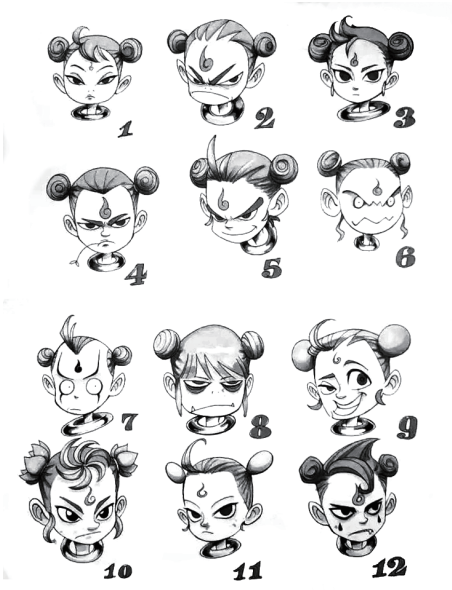
the Base team at Base Media introduced the AI tool

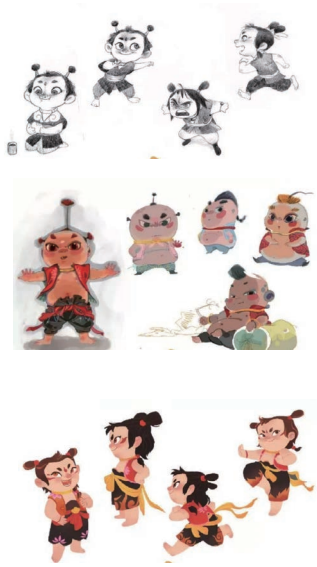
ControlNet:Creation time from weeks to **3** days

iterating **1200** concept images.Generating over **800+**

versions of Nezha's image.The final "Demon Child Nezha" image

achieving a **92.3%** audience satisfaction rate.





Voice Recording

In traditional production modes, directors, screenwriters, animators, editors, and all voice actors need to conduct multiple rehearsals and formal recordings. Now, large models can generate voice tones and styles that match the character's appearance, actions, and expressions based on concept images.

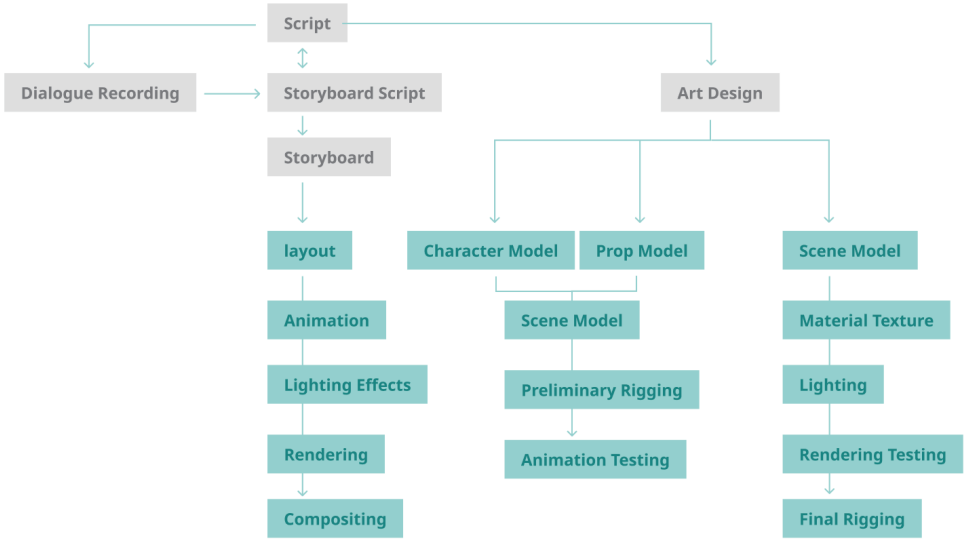
Storyboard Production

Storyboard production is the process by which the director breaks down the script into segments and shots. Before creation, the final project's delivery standards need to be determined, including aspect ratios. If there is a video model that can adapt to any aspect ratio, creators do not need to be concerned about aspect ratio issues.

Production

As the core stage of animation production, the production stage mainly includes asset creation, lighting, special effects, and compositing.

The creation of 3D animation assets is based on pre-production settings, involving asset modeling, texturing, rigging, electronic storyboard layout, and animation. The following focuses on the layout and animation stages.



Layout Stage

Multimodal Generative Technology can quickly generate basic animation elements or scene frameworks based on input creator settings, providing references for layout design. The scenes of the Heavenly Palace and Dragon Palace in "Nezha 2" are stunning and packed with breathtaking fantasy. The construction of these complex scenes also relies on AI technology.

The AI scene generator can automatically generate building structures and texture maps based on input keywords, such as "underwater dragon palace + cyberpunk."

The ice crystal throne of Ao Bing was initially generated by AI in 37 versions, enabling the art director to piece together the final effect by sliding their finger on the touch screen, like playing "Minecraft." This AI-assisted scene generation method greatly saves the art team's time and effort while providing more creative inspiration for creators.

Animation Production

After the layout is approved, the final asset rigging and animation production stages begin. Combining the script, storyboard content, character personalities, and story expression, the actions, expressions, and interactions between assets are adjusted.

This process generally includes three steps: Blocking, Animation, and Facial.

Step	Specific Actions	Traditional Mode	AIGC Mode
Blocking	Setting key actions	Manually drawing original sketches, repeatedly adjusting andmodifying based on personal experience	Automatically capturing complex actions in image information, extracting keyframes, setting, and optimizing key actions
Animation	Inserting in-between frames, refining animation	Adjusting bone positions and rotations frame by frame, manually adjusting animation curves	Automatically filling in the blanks between keyframes, generating coherent and natural content in-between
Facial	Adjusting facial expression details and lip-syncing	Using live-action video as a reference for performance	No need for live-action video reference, adjusting movement amplitude, changing the speed and amplitude of facial muscle movements to affect expression details

During the Blocking stage, "Nezha 2" performed an "INTELLIGENT UPGRADE" of keyframe motion capture, using the image-to-image function of Stable Diffusion.

The original artist only needed to outline the character's sketch, and the AI could automatically fill in colors, lighting details, and generate coherent action sequences.

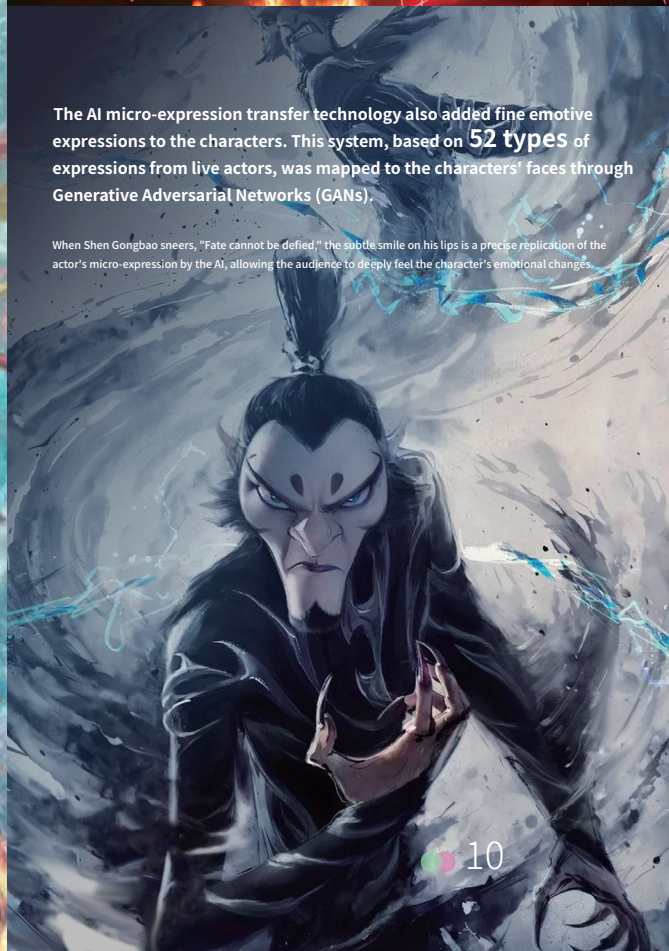
The 1080-degree rotating shot of Nezha on his Wind Fire Wheels was the optimal solution after over 200 iterations by the AI.

During the facial expression adjustment stage, the team established an AI facial expression database, collecting over 100,000 micro-expression data points.

When Nezha says, "My fate is determined by me, not by heaven," the AI automatically matched the combination of pupil contraction and mouth tremor, increasing the accuracy of emotion transmission by 70%.

The AI micro-expression transfer technology also added fine emotive expressions to the characters. This system, based on 52 types of expressions from live actors, was mapped to the characters' faces through Generative Adversarial Networks (GANs).

When Shen Gongbao sneers, "Fate cannot be defied," the subtle smile on his lips is a precise replication of the actor's micro-expression by the AI, allowing the audience to deeply feel the character's emotional changes.



It is worth noting that the production stage, especially the Layout and Animation (keyframe and in-between frame) stages, are the most time-consuming periods of the entire animation production process. In traditional production modes, creating high-quality animation, a 5-minute 2D animation episode's Layout and Animation stages typically take about six months due to frame-by-frame drawing. Although 3D animation involves complex model building and rigging in the early stages, keyframe and in-between frame production have automation advantages with software assistance, generally taking requiring four to six months per episode.

The application of Multimodal Generative Technology is expected to significantly reduce labor costs in these parts.

Stage	Specific Work	2D Animation Episode Time	3D Animation Episode Time	AIGC Time	Cost Reduction
Layout	Determining character and scene positions and composition	2-3 months	1-2 months	10-15 days	45%
Keyframe Animation	Setting key action points like running and landing	1-2 months	1-2 months	15-20 days	38%
In-between Animation	Filling in frames between keyframes for smooth motion	2-3 months	2-3 months	1-1.5 months	50%

Data Source: "2025 China Animation Film Industrialization White Paper"

In "Nezha 2," **30%** of the **1948** special effect shots involved repetitive work handled by AI technology, from hair rendering to particle effects. AI algorithms reduced traditional manual time by **60%** and costs by **45%**

Special Effects

According to Base Media, the special effects production efficiency increased by **400%** with single-frame rendering costs dropping from 300 RMB to 18 RMB

Their AI cluster system improved the efficiency of the "Ten Thousand Dragon Armor" scene's special effects by **300%** reducing single-frame rendering time from 72 hours to 8 hours and the overall production cycle from 3 months to 2 weeks with a **40%** cost reduction

Post-production

After animation production is completed, it enters the post-production stage, including voice processing, color grading, subtitles, and other editing processes. The post-production process mainly consists of sound processing and image processing.

Sound Processing

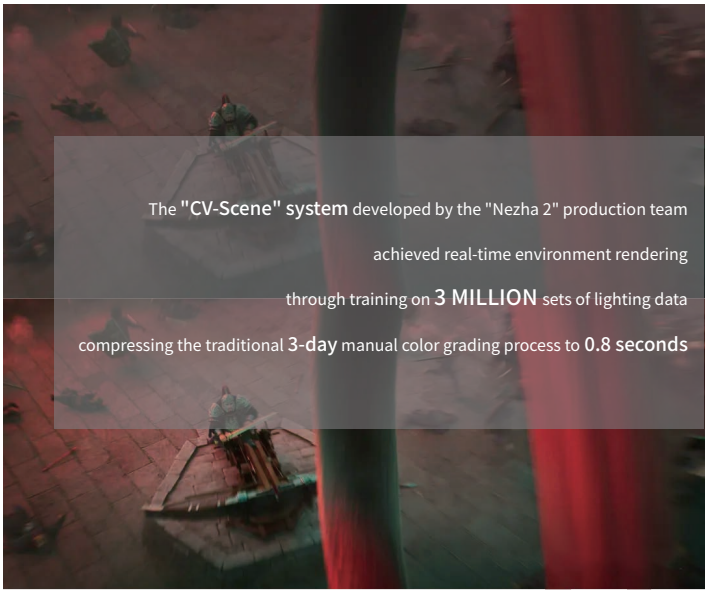
This includes music and sound (dialogue, mimic, sound effects), with mixing and modification checks at each stage and for the same elements, ultimately mixing the processed material with the image. Multimodal Generative Technology can simultaneously process image and audio information, generating matching sound effects and music reference materials, effectively simplifying this tedious work.

AI voice synthesis technology generated seven dialect versions of the voice for "Nezha 2," boosting the regional market box office by 15%-20%.

Image Processing

Color correction is crucial yet repetitive, including adjustments to contrast, tone, color, brightness, saturation, and density, aiming to unify the color of each shot and ensure the work's tonal harmony.

Multimodal Generative Technology not only ensures color consistency between different frames but also makes color presentation harmonious and natural, greatly reducing reliance on manual color grading.



02 ESTABLISHING EVALUATION STANDARDS FOR VIDEO MODELS BASED ON ANIMATION APPLICATION SCENARIOS

“Based on in-depth practice in different application scenarios throughout the full process, CreateAI has constructed a comprehensive and targeted video model evaluation system from the actual needs of animation production.”

Application Scenario Needs	Dimension/Players
Generating high-quality visual art materials in pre-production, automatic in-between frame generation meeting the same standards	Quality
Controlling generated angles, actions, movement amplitude, and speed according to the requirements of the layout's visual language	Controllability
High consistency in generating characters, scenes, props, etc., when generating keyframes and automatically filling in-between frames	Consistency
Generating diverse animation styles such as Ghibli's naturalism, Pixar's realism, and Disney's fairy tale aesthetics based on 2D/3D technology	Style
Significant cost advantage compared to manual labor	Cost
Animation creators are mostly small studios	Usability

This system includes six key dimensions, aiming to accurately measure the performance of video generation models in animation creation tasks, ensuring that the models can truly meet the strict requirements of various stages of the animation industry.

03 INDUSTRIAL-GRADE IMAGE-TO-VIDEO MODEL "RUYI" - CUSTOM DEVELOPMENT TO BALANCE SCENARIO NEEDS

BACKGROUND OF CUSTOM DEVELOPMENT

In the highly competitive field of creative content creation, general-purpose foundational models face many challenges. From one aspect, inconsistent training data leads to significant differences in generative effects for different themes. Post-adjustment and correction are difficult, making it hard to precisely meet diverse creative needs. On the other hand, lack of or inaccurate data annotation leads to deviations in the model's understanding and generation of content, failing to meet creators' requirements for detail and accuracy. To address these pain points, CreateAI decided to independently develop the industrial-grade image-to-video model, called "Ruyi" to ensure deep alignment with the animation market's demand.

DATA-LEVEL OPTIMIZATION

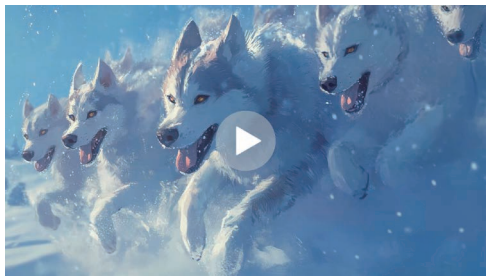
The Ruyi image-to-video model has been meticulously designed in data collection and processing. The team extensively collected data covering diverse themes such as animation, science fiction, and live-action, ensuring balanced data distribution through scientific screening and matching. At the same time, leveraging global leading data annotation experience and capabilities, the team adopted a combination of automated and manual annotation to ensure high-quality data annotation. Balanced and high-quality data thus makes Ruyi more stable and superior regarding various creative needs, being friendlier to creators.

INDUSTRIAL-GRADE USAGE STANDARDS

Currently, combining user feedback for continuous updates and iterations, Ruyi has achieved industrial-grade usage standards in the following five aspects:

■ Ultra-High-Quality Output

Ruyi's color presentation is harmonious and natural. Its architecture is based on the DiT (Denoising Diffusion Transformer) model, with the Diffusion Transformer module generating videos after data compression, ensuring fine and high-fidelity generative effects through 3D full attention and various positional encoding methods.



CreateAI

Scan QR → Follow CreateAI

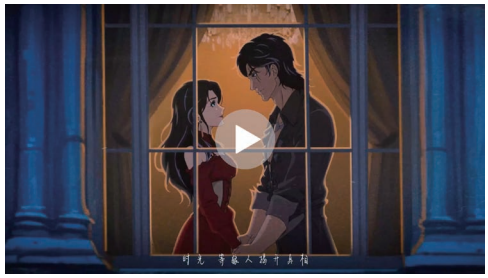
Excellent Consistency and Coherence

Ruyi excels in frame-to-frame consistency and smoothness of action, adopting a frame rate of 24 frames per second, resulting in smoother action.

CreateAI

Scan QR → Follow CreateAI

In terms of cross-frame color consistency control, Ruyi has pioneered the solution to the technical shortcomings present in most models on the market. It can precisely control colors, ensuring that the generated materials maintain consistent background colors throughout.



CreateAI

Scan QR → Follow CreateAI

Highly Controllable Generation

Resolution, Duration, and Aspect Ratio Control : Ruyi supports resolutions from 384×384 to1024×1024, and can generate any aspect ratio, as well as videos up to 120 frames/5 seconds.

CreateAI

Scan QR → Follow CreateAI

Start and End Frame Control : Supports video generation based on up to five start frames and five end frames, allowing creators to control the start and end frames of a video, meeting editing and content planning needs.

CreateAI

Scan QR → Follow CreateAI

Motion Amplitude Control : Provides four levels of motion amplitude control, allowing creators to adjust according to their needs, achieving controllable motion effects.

CreateAI

Scan QR → Follow CreateAI

Camera Control : Provides five camera control options: up, down, left, right, and stationary, allowing creators to control the visual angle and create different visual effects.

CreateAI

Scan QR → Follow CreateAI

■ Efficient Creation Cost Control

Ruyi aims to reduce the development cycle and cost of content. Upon release, the Ruyi-Mini-7B version was open-sourced for creators to use for free. Its DiT model architecture, with the CasualVAE module responsible for video data compression and decompression, greatly improves generation efficiency, allowing creators to complete high-quality video production in a short time, reducing labor and time costs.

■ Convenient Creation Interaction Experience

"Ruyi" is designed for consumer-grade graphics cards (such as RTX 4090), providing detailed deployment instructions and ComfyUI workflows, allowing even beginners to quickly get started, lowering the entry barrier.

04 CHALLENGES AND PROSPECTS OF MULTIMODAL GENERATIVE TECHNOLOGY IN ANIMATION PRODUCTION

CHALLENGES

Data Dependency

The generative effects of Multimodal Generative Technology highly depends on high-quality multi-dimensional data (text/image/audio/video,etc.). Annotation errors, modal loss, or distribution imbalance can easily lead to content distortion. During the art design stage of "Nezha 2," over 800TB of multimodal training data - including text, storyboards, audio, and video - was consumed to achieve film-grade standards in hair rendering and lighting effects.

Poor Physical Likeness

When simulating object motion and spatial interactions, current technology has shortcomings in physical likeness, resulting in some animation scenes not conforming to real-world physical principles, such as abnormal object motion trajectories and unrealistic collision effects, greatly diminishing the realism and immersion of the animation. In "Nezha 2," 12% of the special effects shots required manual correction due to physical distortion, increasing costs by RMB 23 million.

Integration Challenges

Effectively integrating Multimodal Generative Technology with existing animation production processes and tools is not easy, requiring solutions for technical interfaces, data format compatibility, and other issues to ensure seamless integration into current workflows.

FUTURE PROSPECTS

TECHNOLOGICAL BREAKTHROUGHS

Special effects adhering to physical laws will become more realistic and impactful. With the continuous development of deep learning, computer vision, and natural language processing technologies, Multimodal Generative Technology is expected to combine with physical simulation engines (such as Houdini) to more accurately simulate the real world, addressing physical distortion issues.

INDUSTRY TRANSFORMATION

Multimodal Generative Technology will drive changes in the production methods of the animation industry, significantly shortening production cycles and reducing costs, enabling the development of more creative ideas. It will also change the division of labor and collaboration modes in the animation industry, creating new professions and positions. Regarding industry recruitment needs, talent with strength in privacy management, art background, physical foundation, or proficiency in computer languages will be in higher demand.

CONTENT INNOVATION

This technology provides animation creators with more creative possibilities, stimulating creative inspiration and promoting the diversification and personalization of animation content to meet the growing demands of different audience groups.

CONCLUSION

“Multimodal Generative Technology, as an emerging force in the animation production field, has shown great application potential and innovative value. Despite facing many challenges, with continuous technological advancements and improvements, it will play an increasingly important role in the full process of animation production.

According to IDC, by the end of 2025, AIGC will cover 70% of standardized processes in animation production.”

CreateAI officially launched an industrial-grade image-to-video model “Ruyi” on December 17, 2024. We believe that given the current quality and state of industry technology, conditions are ripe for productized technology validation. By the end of 2025, even though manual original production remains the foundation for ensuring production quality, video models will be deeply integrated into the production process of industrial animation series.

In future updates, Ruyi will also have the capability to automatically select the optimal continuation key frames from the generated videos. This will prioritize overcoming existing industry technical bottlenecks and completely address the core challenge in animation production of "easy action continuity, difficult narrative direction." And we will produce animation shorts based on it to visually demonstrate the application effects and advantages of the technology. Indeed, we have set a more long-term goal: to reduce animation production costs and development cycles by 50% within the next five years through continuous technological breakthroughs. Achieving this goal will significantly alleviate the cost pressures on animation production companies, helping more high-quality animation works reach audiences more efficiently.

We look forward to working with industry stakeholders to fully explore the potential of Multimodal Generative Technology and jointly drive the animation industry to reach new heights.

APPENDIX

– Data Source Description

Box Office Data:

Maoyan Professional Edition, Box Office Mojo

Technical Efficiency Data:

Internal data from the production team of "Nezha 2," user testing reports from the film's production team;

Content from mainstream media and professional platforms such as the Chinese Social Sciences Network, CCTV Finance, and ZCOOL.

Sohu:AI Empowering Animation Creation: The Future of Digital Cultural and Creative Industries from "Nezha 2";

Toutiao:Revealing "Nezha: The Demon Child Conquers th e Sea": How AI Technology Helps Chinese Animation Shine Again?

Toutiao:AI Empowering Chinese Animation, Jiangsu Suqian Eight-Dimensional Vocational Education Leading New Trends in Cultural and Creative Industries;

NetEase:"Nezha: The Demon Child Conquers the Sea" Box Office Exceeds 70 Billion, AI Technology Helps Chinese Animation Rise;

Toutiao:AI Nezha Crushes Hollywood with 12.8 Billion! The Truth Behind China's Animation Reversal;
The Real "Invisible Director" Behind the 120 Billion Myth of "Nezha 2";

"Nezha 2" Global Hit, AI Helping "Break the Cauldron";

Industry Comparison Data:

EntGroup

IDC "Global DataSphere Forecast, 2023–2028"

"2025 China Animation Film Industrialization White Paper"

Create AI

CreateAI is redefining the boundaries of what's possible in digital storytelling. Pioneering the future of entertainment content production by blending cutting-edge generative AI technology with the creativity of world-class creators.

共探新境

C R E A T I V I T Y
R E D E F I N E D

Scan QR to follow official account

reply "Report" for PDF