Computer Modeling in Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.066647

REVIEW

Anime Generation through Diffusion and Language Models: A Comprehensive Survey of Techniques and Trends

Yujie Wu[1], Xing Deng[1,*], Haijian Shao[1], Ke Cheng[1], Ming Zhang[1], Yingtao Jiang[2] and Fei Wang[1]

[1]School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang, 212003, China
[2]Department of Electrical and Computer Engineering University of Nevada, Las Vegas, NV 89154, USA

*Corresponding Author: Xing Deng. Email: xdeng@just.edu.cn

Received: 14 April 2025; Accepted: 25 August 2025; Published: 30 September 2025

ABSTRACT: The application of generative artificial intelligence (AI) is bringing about notable changes in anime creation. This paper surveys recent advancements and applications of diffusion and language models in anime generation, focusing on their demonstrated potential to enhance production efficiency through automation and personalization. Despite these benefits, it is crucial to acknowledge the substantial initial computational investments required for training and deploying these models. We conduct an in-depth survey of cutting-edge generative AI technologies, encompassing models such as Stable Diffusion and GPT, and appraise pivotal large-scale datasets alongside quantifiable evaluation metrics. Review of the surveyed literature indicates the achievement of considerable maturity in the capacity of AI models to synthesize high-quality, aesthetically compelling anime visual images from textual prompts, alongside discernible progress in the generation of coherent narratives. However, achieving perfect long-form consistency, mitigating artifacts like flickering in video sequences, and enabling fine-grained artistic control remain critical ongoing challenges. Building upon these advancements, research efforts have increasingly pivoted towards the synthesis of higher-dimensional content, such as video and three-dimensional assets, with recent studies demonstrating significant progress in this burgeoning field. Nevertheless, formidable challenges endure amidst these advancements. Foremost among these are the substantial computational exigencies requisite for training and deploying these sophisticated models, particularly pronounced in the realm of high-dimensional generation such as video synthesis. Additional persistent hurdles include maintaining spatial-temporal consistency across complex scenes and mitigating ethical considerations surrounding bias and the preservation of human creative autonomy. This research underscores the transformative potential and inherent complexities of AI-driven synergy within the creative industries. We posit that future research should be dedicated to the synergistic fusion of diffusion and autoregressive models, the integration of multimodal inputs, and the balanced consideration of ethical implications, particularly regarding bias and the preservation of human creative autonomy, thereby establishing a robust foundation for the advancement of anime creation and the broader landscape of AI-driven content generation.

KEYWORDS: Diffusion models; language models; anime generation; image synthesis; video generation; stable diffusion; AIGC

1 Introduction

At the forefront of this transformation are diffusion models and language models, two classes of generative AI designed to bridge textual and visual domains. These powerful models are significantly impacting certain creative sectors like anime generation and are increasingly relevant across diverse scientific

Copyright © 2025 The Authors. Published by Tech Science Press.
This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

and engineering fields [1]. Diffusion models have demonstrated strong capabilities in synthesizing high-quality images and videos [2,3]. Studies have shown that these models have improved performance improved performance compared to traditional generative adversarial networks (GANs) in image synthesis in terms of general quality, diversity, and specific metrics like FID [4–6]. Meanwhile, Language models such as BERT and GPT have significantly advanced the field of natural language processing [7,8]. The integration of these technologies has enabled systems to interpret textual prompts and generate corresponding anime-style visuals with notable fidelity, a capability demonstrated by models like Stable Diffusion [3].

This survey focuses on pivotal diffusion and language models that have demonstrated substantial impact and are directly relevant to anime content generation. Specifically, this comprehensive review addresses the pivotal question: How are diffusion and language models currently advancing and transforming the landscape of anime content generation, and what are the key challenges and future directions in effectively leveraging these technologies? Our selection criteria prioritize models with demonstrated efficacy in synthesizing anime-style visuals, coherent narratives, and related creative assets, alongside their technical innovation and prominence within the generative AI landscape.

This study investigates the application of these models across key anime production domains: narrative and graphic novel genesis, illustrative and keyframe synthesis, and episodic and interactive media expansion. For instance, language models can automate script generation and dialogue creation, while diffusion models synthesize anime-stylized imagery and sequential frames [9,10]. The integration of these technologies addresses enduring challenges in creative efficiency, content personalization, and fiscal optimization. While necessitating significant computational resources and capital outlay for high-performance infrastructure, these technologies can streamline certain manual workflows, potentially leading to time savings and reduced labor-intensive operational costs in anime production. This shift has led to notable advancements by streamlining complex processes and accelerating content generation in specific areas [11]. The use of large-scale datasets like LAION-5B enhances the multilingual and stylistic capabilities of these models [12], broadening their applicability. Furthermore, a synergistic analysis employing automated (e.g., Fréchet Inception Distance, CLIP Score) metrics alongside human-centered evaluation provides a comprehensive paradigm for appraising the efficacy of sophisticated diffusion and language models. This approach elucidates their technical viability and observed capacity to foster creative innovation within animation production [13,14], thereby elucidating both their technical viability and their capacity to foster creative innovation within animation production.

Despite their promise, these technologies face significant hurdles. Technical challenges include maintaining consistency across multi-frame sequences and multi-character scenes. Ethical considerations, such as originality, copyright, and the potential diminishment of human creativity, also loom large [15].

The overall structure of this study is depicted in Fig. 1: Section 2 offers a comprehensive background on diffusion models and language models, detailing their theoretical foundations and development history. Section 3 explores image generation methodologies, focusing on Stable Diffusion and its ecosystem for anime-style synthesis. Section 4 examines video generation, addressing advancements in temporal consistency and character animation. Section 5 delves into music composition, while Section 6 investigates game generation, extending the application of these models to interactive media. Section 7 covers alternative applications, such as narrative synthesis and virtual streamers. Section 8 discusses the ethical implications and reviews the progress and foundational challenges of generative AI. Section 9 concludes the paper by synthesizing the key findings and proposing directions for further research.

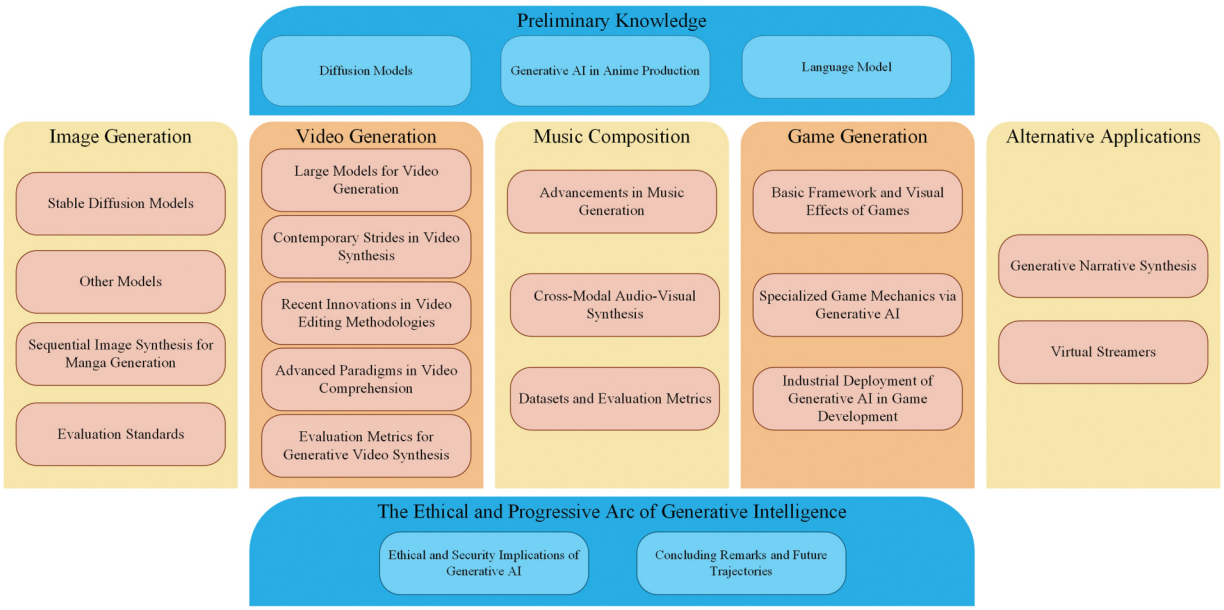


Figure 1: The generative intelligence framework: a structural overview

2 Preliminary Knowledge

2.1 Generative AI in Anime Production

The anime industry is undergoing notable changes precipitated by the growing emergence of Generative Artificial Intelligence (GAI). The synergistic integration of Natural Language Processing (NLP), Computer Vision (CV), and cross-modal synthesis offers a *potential* pathway to automate certain conventional manual workflows. While these advanced models incur high computational overhead, their ability to reduce human effort and accelerate content iteration *can represent* a strategic shift towards more efficient and cost-effective production paradigms. Language Models (LMs) and Diffusion Models have emerged as pivotal instruments, providing the field with enhanced content comprehension and generative capabilities. This significant influence of AI in creative industries necessitates a judicious equilibrium between technological innovation and the preservation of human ingenuity. While AI automates repetitive tasks across numerous sectors, within creative domains like anime, it functions as a collaborative instrument, enabling novel creative avenues, optimizing workflows, and enhancing creative processes [15]. Nevertheless, maintaining the human element and inherent authenticity that define the output of creative industries remains paramount [15].

This significant evolution is impacting various echelons of anime production, from the foundational literary and visual substrates to certain culminating deliverables. Specifically, the following critical domains are undergoing **notable developments**:

Narrative and Graphic Novel Genesis: Literary works, serving as rich repositories of imaginative narratives, frequently constitute the genesis for anime adaptations, while graphic novels (manga) represent a sophisticated synthesis of visual and textual storytelling. AI's role encompasses facilitating script generation and potentially influencing narrative architectures.

Illustrative and Keyframe Synthesis: Illustrations, conveying nuanced emotional expression and visual narratives, and key animation, defining pivotal movement frames, are critical constituents of the animation process. AI is being leveraged to expedite the colorization of anime line drawings [9], synthesize anime-stylized imagery (Yang) [10], and even contribute to character conceptualization (Tang and Chen) [16].

Episodic and Interactive Media Expansion: Anime television series, a cornerstone of the industry and a primary revenue stream, and interactive media, notably role-playing games (RPGs), amplify the influence of anime narratives through immersive engagement. AI contributes to enhanced efficiency across various stages of animation production, including pre-production, asset creation, animation production, and post-production [17].

The confluence of sophisticated LMs, such as GPT and BERT, which exhibit aptitude in generating coherent scripts and dialogues, with advanced Diffusion Models, adept at synthesizing anime-stylized visuals, offers *potential* solutions to enduring challenges pertaining to creative efficiency, content personalization, and fiscal optimization. This synergistic integration aims to bridge textual narratives with visual content, contributing to a period of notable advancement in certain aspects of the anime creation paradigm. This integration is facilitating *some aspects of* industrial upgrading, fostering innovation, and augmenting productivity within *parts of* the digital creative industry (Wagan and Sidra) [11].

2.2 Diffusion Models

Diffusion models [3], a class of generative models, leverage the principle of stochastic reverse diffusion to synthesize images from latent noise. The core mechanistic paradigm involves a forward diffusion process, iteratively corrupting data with Gaussian noise, followed by a reverse denoising process, reconstructing the image from the noise distribution. In essence, these models learn to invert the progressive degradation of data structure, enabling the recovery of high-fidelity images. This process is illustrated in Fig. 2.

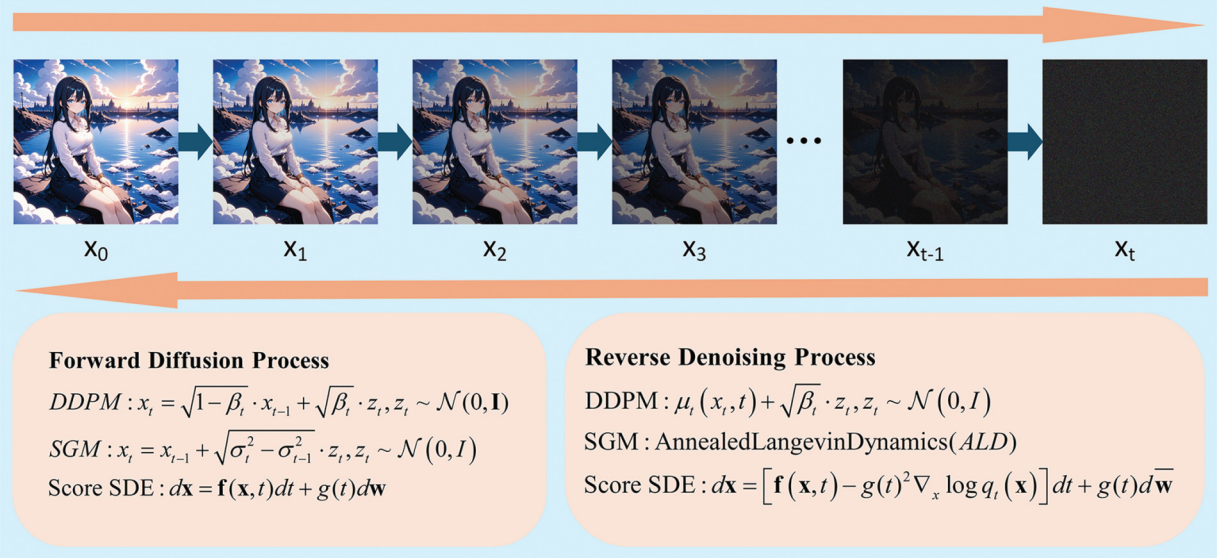


Figure 2: Overview of diffusion models (DDPM, SGM, and Score SDE diffusion and denoising processes)

2.2.1 Mathematical Formalization of Diffusion Processes

Diffusion models operate on the principle of incrementally transforming data through a **forward diffusion process** and subsequently reversing this transformation via a **denoising process** to generate novel samples.

The **forward process** systematically introduces Gaussian noise into an initial image x_0 across **T** discrete timesteps, progressively corrupting it until it approximates a pure noise distribution x_T. This controlled degradation is governed by a predetermined noise schedule.

Conversely, the reverse denoising process endeavors to iteratively reconstruct the original image from noise. This is achieved by training a neural network to predict the subtle noise component at each timestep, effectively learning to reverse the corruption introduced by the forward process. The core objective during training is to parameterize this denoising network, enabling it to accurately approximate the conditional probability distribution of a slightly less noisy image given its noisy counterpart.

Generative sampling leverages this trained denoising network. It commences with a random sample drawn from a Gaussian noise distribution, analogous to x_T. The network then iteratively refines this noisy input through successive denoising steps, gradually transforming the pure noise into a coherent, synthesized image x_0.

The optimization of diffusion models primarily involves minimizing the Evidence Lower Bound (ELBO). This objective function quantifies the discrepancy between the forward diffusion process and the model's learned reverse denoising capabilities, effectively guiding the network to accurately reverse the noise corruption.

For readers seeking a comprehensive mathematical treatment of these processes, we refer to the foundational works cited in the original text [18–21].

2.2.2 Development History of Diffusion Models

Fig. 3 shows the historical development of diffusion models. The trajectory of diffusion models has been marked by distinct phases, each characterized by pivotal advancements that have collectively propelled their capabilities from theoretical constructs to increasingly powerful generative tools, influencing fields like anime generation.

The initial phase, ignited by the advent of Denoising Diffusion Probabilistic Models (DDPMs), focused on solidifying theoretical foundations [2]. Key innovations like Denoising Diffusion Implicit Models (DDIMs) significantly enhanced sampling efficiency, substantially reducing the steps required for high-fidelity generation [22]. Concurrently, Classifier Guidance and its successor, Classifier-Free Guidance, markedly improved conditional image synthesis, laying the groundwork for the sophisticated text-to-image models that followed [5,23]. Theoretical explorations into discrete diffusion models, exemplified by Multinomial Diffusion and D3PM, also contributed to this foundational period [23,24].

Subsequent developments centered on diversifying applications and enhancing scalability. Techniques such as Latent Diffusion and VQ Diffusion were instrumental in applying diffusion models to large-scale datasets, making them amenable to real-world applications [3,25]. This period also saw significant strides in sampling acceleration, with methods like PNDM and Analytic-DPM that further reduced computational overhead while maintaining generative quality [26,27]. The observed versatility of these models has extended their utility beyond mere image generation to tasks like semantic segmentation and advanced image editing.

The focus then shifted toward the creation of large-scale models, particularly in the text-to-image domain. Influential models like DALLE-2 and Imagen showcased impressive capabilities in synthesizing images from textual prompts, leveraging vast datasets [28,29]. The release of open-source initiatives like Stable Diffusion, alongside accompanying massive datasets such as Laion-5B, has made these powerful generative tools widely accessible [3,12]. This has led to significant and rapid adoption within research communities, among individual creators, and in agile development environments, fostering experimentation and innovation in AI-driven content generation.

This widespread adoption in research and creative exploration has rapidly led to a phase of deployed applications and domain expansion, manifesting primarily in specialized tools, academic prototypes, and niche creative workflows, rather than comprehensive industry-wide overhauls. Stable Diffusion, in particular,

became a cornerstone for diverse applications [3], including image inpainting (e.g., Equilibrium Diffusion, Shadow Diffusion) [30,31], image perception, 3D generation (e.g., DreamFusion, Magic3D) [32,33], video generation (e.g., Latent Video Diffusion) [34], and medical imaging (e.g., MedSegDiff) [35]. This marked a crucial transition from academic research to practical utility.

The most recent phase emphasizes controllability and cross-domain innovation. Tools like ControlNet have enabled precise manipulation of generated images through explicit conditions (e.g., edge maps, depth maps), offering enhanced creative control. Advancements in text-to-3D generation (e.g., Point-E, DreamFusion), and text-driven video synthesis (e.g., Video Diffusion Models, Make-A-Video) further extended their capabilities [32,36–38]. Concurrently, ongoing research into computational efficiency (e.g., Latent Diffusion [3], Efficient Diffusion) and multi-modal fusion (e.g., Diffusion-LM) continues to explore and enhance the capabilities of diffusion models, integrating disparate data types for more complex and nuanced generative tasks, including their observed impact on anime generation through specialized applications and stylistic control [39,40].



Figure 3: Historical development of diffusion models

2.2.3 Datasets for Diffusion Model Training

The training of diffusion models, irrespective of modality (image, video, audio), necessitates large-scale, high-fidelity datasets [41]. Optimal datasets are characterized by: **Extensive Cardinality:** Datasets comprising hundreds of millions to billions of paired data samples (e.g., image-text) are requisite for capturing the inherent diversity and complexity of the data manifold [42]. **Comprehensive Heterogeneity:** Datasets must encompass a broad spectrum of scenes, styles, and linguistic representations to ensure

robust generalization of generated outputs [43]. **Precise Annotation Fidelity:** Particularly for conditional generative tasks, the semantic coherence between textual and visual/temporal data is paramount. Exemplary datasets are summarized in Table 1.

Exemplary Datasets for Diffusion Model Training:

LAION-5B (March 2022) [12]: A corpus of 5.85 billion image-text pairs, curated via CLIP-based filtering, exhibiting multilingual scope. This dataset has facilitated the training of models such as Stable Diffusion, enhancing generative fidelity and zero-shot capabilities.

Re-LAION-5B (August 2024) [44]: An augmented iteration of LAION-5B, incorporating stringent filtering to mitigate illicit content and providing academically compliant subsets, thereby addressing ethical and legal considerations.

WebVid-10M (March 2021) [45]: A video dataset comprising 10.7 million video clips (52,000 h) with alt-text annotations, contributing to improved temporal consistency and zero-shot video generation.

HD-Vila-100M (2022) [46]: A large-scale video dataset consisting of 100 million high-definition videos (371,000 h) with automatically transcribed textual data, supporting generalized video synthesis tasks.

VidProM (March 2024) [47]: A synthetic video dataset featuring 6.69 million generated video clips (1.6-3 s each), synthesized using multiple generative models, and augmented with NSFW detection and prompt embeddings, designed to facilitate model evaluation and prompt engineering research.

Table 1: Datasets for diffusion models

Dataset name	Release date	Sample size	Primary use	Characteristics
LAION-5B [12]	2022-03	5.85 Billion image-text Pairs	Text-to-image generation, multilingual research	Multilingual, large-scale
Re-LAION-5B [44]	2024-08	5.85 Billion (Updated)	Safety research, language-visual learning	Enhanced safety
ImageNet [48]	2009	1.2 Million images	Unconditional image generation, classification	Hierarchical classification structure
COCO [49]	2017	120,000 images	Scene generation, object detection	Rich annotations
CelebA [50]	2015	203,000 Images	Face generation, attribute recognition	40 facial attribute annotations, aligned faces
WebVid-10M [45]	2021-03	10.7 Million video clips	Text-to-video generation	High diversity
HD-Vila-100M [46]	2022	100 Million video clips	General video generation	High definition, large-scale
VidProM [47]	2024-03	6.69 Million videos	Prompt evaluation, model comparison	Novel, evaluation-oriented

2.2.4 Architectural Advantages

The architectural design of Diffusion Models offers several advantages in generative tasks, notably in latent space operation, flexibility and tractability, neural network adaptability, conditional generation capabilities, and scalability. Many Diffusion Models, such as Stable Diffusion, employ a two-stage training paradigm that first compresses high-dimensional image data into a lower-dimensional latent space via an autoencoder [44]. This approach not only substantially reduces computational complexity but also enhances model scalability, particularly for high-resolution image synthesis. By conducting diffusion and reverse diffusion within this latent space, models like Latent Diffusion Models (LDMs) can efficiently process intricate data while often preserving high-quality generative outcomes [3]. Diffusion Models can effectively balance the analytical tractability of simpler distributions (e.g., Gaussian) with the expressive power of complex models (e.g., GANs), enabling them to model sophisticated data distributions with notable training

stability and sampling efficiency, and have often outperformed traditional generative models as evidenced by their superiority over GANs in image synthesis [51]. The reverse diffusion process is typically orchestrated by flexible neural network architectures, including U-Net or Transformer variants, allowing for task-specific customization; for instance, Stable Diffusion leverages U-Net with cross-attention, while Stable Diffusion 3 employs a Diffusion Transformer (DiT), showcasing this architectural versatility [52,53]. Furthermore, Diffusion Models excel in conditional generation through mechanisms like cross-attention or dedicated conditioning modules (e.g., text encoders), enabling models such as Stable Diffusion and DALL-E 2 to synthesize imagery from textual prompts, thereby broadening their applicability to tasks like text-to-image and layout-to-image generation [28]. Their inherent scalability, facilitated by latent space representations and efficient architectures, allows them to manage high-dimensional data, balancing generative quality with computational expediency; Cascade Diffusion Models, for example, have demonstrated enhanced high-resolution image generation through multi-stage diffusion processes [54].

2.2.5 Training Methodology

The training regimen for Diffusion Models encompasses several critical steps, primarily involving the forward and reverse diffusion processes, the strategic selection of variance schedules, and the application of advanced optimization techniques. The training commences with the forward diffusion process, wherein original data is progressively corrupted by Gaussian noise, transitioning from the data distribution to a pure noise distribution. This process, governed by a predefined variance schedule (typically linear or cosine), functions as a Markov chain, incrementally introducing noise. The judicious choice of variance scheduling significantly impacts training stability and generative quality. Subsequently, the reverse diffusion process trains a neural network to invert this corruption, iteratively denoising from pure noise to reconstruct the original data. The neural network typically predicts the noise or the denoised data at each step, optimizing through the minimization of a simple mean squared error (MSE) based loss function. Variance schedules, which dictate the rate of noise addition, are paramount, with linear and cosine schedules being common choices. Advanced training techniques, such as DDIM (Denoising Diffusion Implicit Models), have markedly accelerated sampling by reducing the number of necessary steps. Progressive Distillation further refines model performance using a teacher-student framework, and Consistency Models enhance generative quality by enforcing specific consistency constraints [2].

Table 2: Historical development of large language models

Year	Model name	Open/closed source	Model type	Dev. Team/Co.
2018	BERT [7]	Open source	Encoder	Google
2019	GPT-2 [55]	Closed source	Decoder	OpenAI
2020	RoBERTa [56]	Open source	Encoder	Facebook
2020	T5 [57]	Open source	Enc-Dec	Google
2020	GPT-3 [58]	Closed source	Decoder	OpenAI
2021	GPT-Neo [59]	Open source	Decoder	EleutherAI
2022	LaMDA [60]	Closed source	Decoder	Google
2022	PaLM [61]	Closed source	Decoder	Google
2022	GPT-3 [58]	Closed source	Decoder	OpenAI
2023	GPT-4 [62]	Closed source	Decoder	OpenAI
2023	LLaMA [63]	Open source	Decoder	Meta

(Continued)

Table 2 (continued)

Year	Model name	Open/closed source	Model type	Dev. Team/Co.
2023	Bard	Closed source	Decoder	Google
2023	Claude	Closed source	Decoder	Anthropic
2024	Gemma 2	Open source	Decoder	Google
2024	Llama 3.1	Open source	Decoder	Meta
2024	OpenAI o1	Closed source	Decoder	OpenAI
2024	Phi 3.5	Open source	Decoder	Microsoft
2024	Qwen2.5	Open source	Decoder	Alibaba Cloud
2025	GPT-4.5	Closed source	Decoder	OpenAI
2025	DeepSeek R1	Open source	Decoder	DeepSeek
2025	Gemini 2.0	Closed source	Decoder	GoogleDeepMind
2025	Grok-3	Closed source	Decoder	xAI
2025	Claude 3.7 Sonnet	Closed source	Decoder	Anthropic

2.3 Language Model

2.3.1 Foundational Architectures of Transformer-Based Language Models

The historical development of these models is summarized in Table 2. Transformer-based language models, fundamental to advancements in natural language processing, are broadly categorized into three core architectural paradigms, each optimized for distinct linguistic tasks [64]:

Encoder-Centric Models: Exemplified by models such as BERT, these architectures are primarily designed for comprehension-oriented tasks [7]. They excel at understanding the intricate semantic relationships within input text by learning deep contextual representations. This makes them highly effective for applications like text classification, named entity recognition, and reading comprehension, where the goal is to extract meaning from existing text.

Encoder-Decoder Hybrid Models: Represented by models like T5, these architectures combine the strengths of both encoders and decoders [57,65]. The encoder processes the input sequence, capturing its semantic essence, while the decoder then uses this understanding to synthesize an output sequence. This dual mechanism makes them highly adept at sequence transduction tasks, such as machine translation and text summarization, where one sequence is transformed into another.

Decoder-Dominant Models: Typified by the GPT series (e.g., GPT-4, LLaMA), these models are built for generative tasks [62,63]. Operating autoregressively, they predict subsequent tokens based on preceding ones, enabling them to produce fluent and contextually coherent text. Their inherent design makes them ideal for applications requiring creative text generation, dialogue systems, and content creation, including potential applications in anime script and narrative generation.

2.3.2 Encoder-Centric Models: Contextual Representation Learning

Encoder-centric models, notably **BERT** [7], are specifically engineered to grasp contextual information within text. Their pre-training focuses on learning deep relationships between words in a given context, making them highly effective at tasks that require understanding existing text, such as text classification

or question-answering. However, their architecture, being primarily focused on comprehension, inherently limits their direct application in generating novel text or dialogues.

2.3.3 Encoder-Decoder Hybrid Models: Sequence Transduction

The architecture of an encoder-decoder model is shown in Fig. 4. Encoder-decoder models facilitate sequence transduction by employing a two-part system. An encoder first processes an input sequence to distill its underlying meaning into a rich contextual representation. This representation is then passed to a decoder, which uses this understanding to construct a new, coherent output sequence. This architecture is highly effective for tasks where the input needs to be transformed into a different output format, such as translating a script from one language to another or summarizing a long narrative into a concise synopsis, which could be valuable for managing anime production content [57,65].

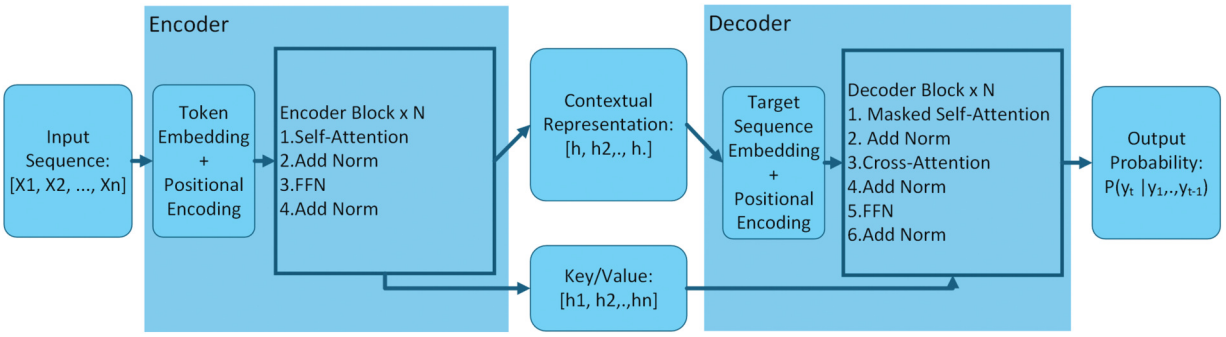


Figure 4: Encoder-decoder model architecture diagram

2.3.4 Decoder-Dominant Models: Autoregressive Sequence Generation

The schematic for a decoder-only model is depicted in Fig. 5. Decoder-dominant models operate autoregressively, generating sequences token-by-token, conditioned on preceding tokens. Input sequences are embedded and processed through stacked decoder blocks, each comprising masked self-attention, add-and-norm, and FFN layers. The output is a probability distribution over subsequent tokens.

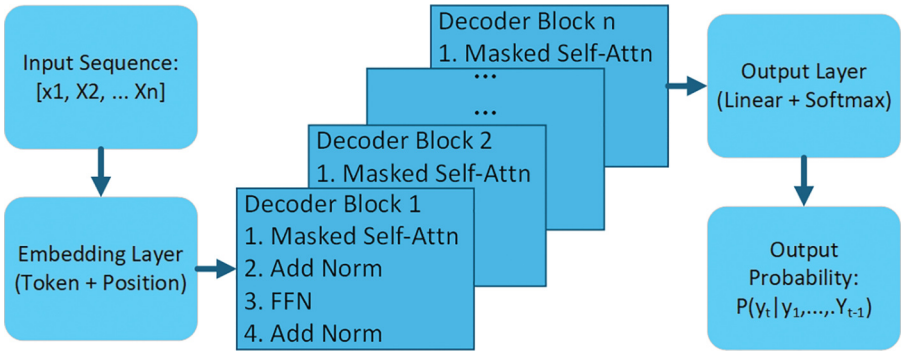


Figure 5: Decoder-only model architecture schematic

The GPT and LLaMA series are representative decoder-dominant models [62,63]. GPT models demonstrate strong generative capabilities, with GPT-4 extending to multimodal content. LLaMA models offer scalable architectures for diverse applications.

2.3.5 Architectural Advantages

The Transformer architecture derives its notable generative capabilities from several key innovations, including the self-attention mechanism, inherent parallelizability, scalability, the absence of recurrent structures, multi-head attention, positional encodings, and a continuous evolution of efficient variants. The self-attention mechanism, a cornerstone of the Transformer, dynamically allocates attention weights based on the relevance of input sequence elements, thereby capturing long-range dependencies without relying on sequential processing, a significant advancement over traditional models in complex tasks like machine translation and text generation [64]. Unlike recurrent neural networks (RNNs), Transformers inherently support parallel processing of entire input sequences, dramatically accelerating training and inference, particularly on GPUs [64]. This parallelization, combined with the ability to handle arbitrary sequence lengths, provides Transformers with high scalability across diverse tasks, enabling models such as BERT and GPT to adapt from text classification to generation via pre-training and fine-tuning [7,66]. The elimination of recurrent structures mitigates the vanishing gradient problem, facilitating the training of deeper, more stable, and efficient networks for long sequences. Multi-head attention further augments representational capacity by allowing the model to concurrently attend to distinct subspaces within the input sequence, instrumental in capturing bidirectional context (e.g., BERT) or facilitating autoregressive generation (e.g., GPT) [7,66]. Given the inherent order-agnostic nature of self-attention, Transformers incorporate positional encodings, such as sinusoidal functions, to convey token position; recent innovations like Rotary Positional Embeddings (RoPE) and ALiBi optimize for relative positional dependencies and enable fine-tuning on longer sequences after pre-training on shorter ones [67,68]. To address the computational demands of the Transformer, numerous efficient variants have emerged, including Reformer, which leverages locality-sensitive hashing (LSH) to reduce attention complexity from $O(N^2)$ to O(NlogN), and BigBird, achieving O(N) complexity via small-world networks [69,70]. Furthermore, FlashAttention and FlashAttention-2 have dramatically accelerated attention computations, reaching speeds up to 230 TFLOPs/s on A100 GPUs [71]. Contemporary trends in Transformer architecture, as of 2025, focus on sparsity through sparse attention mechanisms for reduced computation and improved memory efficiency, Mixture-of-Experts (MoE) models to enhance scalability and efficiency by partitioning the network into specialized modules, and adaptive computation techniques that dynamically adjust computational resources based on input complexity to optimize performance.

2.3.6 Training Methodology

Transformer training typically unfolds in a two-stage paradigm: pre-training and fine-tuning. Initial pre-training occurs on vast corpora (e.g., Wikipedia) using self-supervised objectives such as masked language modeling (BERT) or auto-regressive language modeling (GPT). This is followed by fine-tuning on smaller, task-specific datasets to adapt the model for particular applications, a methodology that significantly curtails training costs and enhances generalization. The standard Transformer architecture comprises an encoder and a decoder. The encoder processes input sequences via multi-head self-attention to generate contextual representations, while the decoder, employing masked self-attention and encoder-decoder attention, produces output sequences auto-regressively. Attention mechanisms are pivotal in training: encoder self-attention captures intra-input relationships; decoder masked self-attention ensures predictions are solely based on preceding tokens; and encoder-decoder attention integrates the encoder's output into the decoder, thereby enhancing generative quality. Layer Normalization and Residual Connections, applied after each sub-layer, are crucial for mitigating the vanishing gradient problem and facilitating the training of deeper networks. Optimization for Transformer training typically employs the Adam optimizer coupled with learning rate schedules (e.g., warmup and decay) to improve convergence, while regularization techniques like Dropout prevent overfitting, especially in models with extensive parameter counts. Recent advancements

in training methodology include efficient inference techniques such as Key-Value Caching to obviate redundant computations of key and value vectors, speculative decoding, and multi-token prediction to balance accuracy and speed in real-time applications. Pre-layer normalization (Pre-LN) has been introduced to enhance training stability. Moreover, the Transformer architecture has been extended for multimodal training, exemplified by models like DALL-E, which jointly process complex datasets encompassing both text and images [64].

2.4 Model Fine-Tuning Techniques and Anime-Related Datasets

2.4.1 Model Fine-Tuning Techniques

An overview of common fine-tuning techniques is provided in Table 3. Full Fine-Tuning, the most direct method, updates all model parameters. While offering maximal adaptation potential for a specific style, it incurs high computational expense, requires substantial data, and exhibits limited generalizability to diverse, unseen anime styles, risking overfitting on limited datasets [72].

Table 3: Overview of common model fine-tuning techniques

Technique name	Primary model type	Description	Modifies original model weights?	Parameter efficiency
LoRA (Low-Rank Adaptation)	LLMs, diffusion models	Freezes pre-trained model weights, injects trainable low-rank decomposition matrices, significantly reducing trainable parameters.	No	High
ControlNet	Diffusion models	Adds additional conditioning (e.g., edges, poses) to diffusion models to control image generation, providing precise guidance.	No	Medium
DreamBooth	Diffusion models	Fine-tunes a diffusion model using a small number of images of a specific subject to generate new images of that subject in different scenarios.	Yes	Low
Textual inversion	Diffusion models	Learns new embeddings for specific visual concepts, enabling diffusion models to generate images based on these concepts.	No	High
HyperNetworks	Diffusion models, LLMs	A neural network generates the weights of another neural network, enabling efficient and flexible model adaptation.	Yes	High (Potential)
Full fine-tuning	LLMs, diffusion models	Trains all parameters of a pre-trained model on a new dataset to adapt to a specific task, incurring high computational cost.	Yes	Low
Feature-based fine-tuning	LLMs	Fine-tunes only the later layers of a pre-trained model, keeping earlier layers frozen, preserving general features for task adaptation.	Partial	Medium
Adapters	LLMs	Adds small task-specific modules to large language models, enabling parameter-efficient fine-tuning.	No	High
Prompt tuning	LLMs	Guides large language model behavior by adjusting input prompts without modifying model weights, offering lightweight adaptation.	No	Extremely high
Prefix-tuning	LLMs	Adds small trainable modules (prefixes) before the input sequence, enabling task-specific adaptation while keeping the model frozen.	No	High
Representation Fine-Tuning (ReFT)	LLMs	Learns task-specific interventions on hidden representations instead of updating weights, achieving parameter-efficient fine-tuning.	No	Extremely high

(Continued)

Table 3 (continued)

Technique name	Primary model type	Description	Modifies original model weights?	Parameter efficiency
Instruction fine-tuning	LLMs	Fine-tunes large language models using instruction-response pairs to enhance their ability to follow instructions and generate task-specific responses.	Yes	Medium
Multi-task fine-tuning	LLMs	Simultaneously fine-tunes a single model on multiple related tasks, learning shared representations and improving performance across tasks.	Yes	Medium
Domain-specific fine-tuning	LLMs, diffusion models	Further trains a pre-trained model on a dataset specific to a particular domain, enhancing performance on tasks within that domain.	Yes	Medium

To mitigate these costs, Parameter-Efficient Fine-Tuning (PEFT) methods update only a small subset of parameters or introduce minimal new ones. LoRA injects low-rank matrices, achieving high efficiency but struggling with blending or switching between vastly different styles [73]. Adapters insert small modules, offering modularity and efficiency at the cost of potential inference latency and configuration complexity; combining style-specific Adapters for generalization is challenging [74]. Prompt Tuning and Prefix-Tuning manipulate input embeddings for extreme efficiency but provide limited control over complex stylistic details [75,76]. Representation Fine-Tuning (ReFT) intervenes on latent representations, offering low cost and non-invasiveness, though its efficacy for complex anime style generalization is under investigation [77].

Diffusion-specific techniques provide conditional control. ControlNet adds spatial conditioning (e.g., edges), potent for structure but relying on the base model's style aptitude [78]. DreamBooth specializes models on specific subjects from few images, achieving high fidelity but overfitting and limiting subject generalization across varied anime styles [79]. Textual Inversion learns new embeddings for concepts, offering efficiency but limited capacity for intricate style representation [80]. HyperNetworks dynamically generate model parameters, promising flexibility but facing training instability and potentially yielding lower quality outputs than direct fine-tuning [81]. Feature-Based Fine-Tuning updates only later layers, computationally lighter but potentially insufficient for style changes requiring lower-level feature modification and bounded in its generalization to diverse styles [82].

Finally, Instruction Fine-Tuning, Multi-Task Fine-Tuning, and Domain-Specific Fine-Tuning enhance task-specific or domain-confined performance. While improving utility within a prescribed scope (e.g., sci-fi anime), they face constraints in achieving comprehensive generalization or synthesizing novel anime styles due to dataset limitations, task interference, and inherent model capacity [83].

While fine-tuning is crucial for anime style acquisition, achieving seamless, high-fidelity generalization across the diverse spectrum of anime aesthetics remains a significant research challenge.

Table 4: Overview of key anime-related datasets

Year	Conference /Journal	Title	Scale
2025	COLING	Context-informed machine translation of manga using multimodal large language models [84]	Constructed a novel evaluation corpus, representing the premiere Japanese-Polish parallel corpus specifically curated for manga translation.
2024	Arxiv	Tails Tell Tales: chapter-wide manga transcriptions with character names [85]	Augments the extant PopManga dataset and established a novel character repository featuring over 11,000 distinct characters.
2024	Arxiv	CoMix: a comprehensive benchmark for multi-task comic understanding [86]	Comprises 3800 images sampled from 100 distinct manga volumes, meticulously annotated with 130,000 object instances, 30,000 text-character linkages, and 33,000 character clusters, encompassing 16,000 unique names.
2023	CVPR	Human-Art: a versatile human-centric dataset bridging natural and artificial scenes [87]	Features 50,000 high-fidelity images, exceeding 123,000 human instances, and spanning 20 diverse scene categories.
2023	Arxiv	Manga109Dialog: a large-scale dialogue dataset for comics speaker detection [88]	Comprises 132,692 speaker-text pairs and 9904 images featuring speaker-text pair annotations. Represents the world's largest dialogue corpus specifically curated for manga speaker detection.
2023	ACM-TG	Semi-supervised reference-based sketch extraction using a contrastive learning framework [89]	Curated a novel dataset of authentic sketches rendered by professional artists, encompassing four distinct sketching styles, each style paired with 25 corresponding color images.
2023	ACM-TG	Parsing-Conditioned Anime Translation: A New Dataset and Method [90]	Introduced the Danbooru-Parsing dataset, featuring 4921 images with dense annotations.
2022	NeurIPS-DB	AnimeRun: 2D animation visual correspondence from open source 3D Movies [91]	A novel corpus for establishing 2D animation visual correspondence.
2022	ECCV	COO: comic onomatopoeia dataset for recognizing arbitrary or truncated texts [92]	Comprises 10,602 images derived from 109 manga volumes, featuring 61,465 polygon annotations and 2261 linkages.
2022	ECCV	AnimeCeleb: large-scale animation celebheads dataset for head reenactment [93]	A substantial-scale corpus of animation character heads specifically designed for head reenactment tasks.
2021	Arxiv	DAF:RE: a challenging, crowd-sourced, large-scale, long-tailed dataseT For Anime Character Recognition [94]	Comprises nearly 500,000 images, spanning over 3000 distinct categories. Represents a large-scale, crowd-sourced, long-tailed benchmark dataset.
2020	ACM-MM	Cartoon face recognition: a benchmark dataset [95]	The iCartoonFace dataset comprises 389,678 images and features 5013 distinct cartoon characters for recognition tasks; it also includes 60,000 images with annotations for 109,810 cartoon faces for detection tasks. At the time of its release, it constituted the largest benchmark dataset for cartoon face recognition.
2020	ECCV Workshop	Unconstrained text detection in manga: a new dataset and baseline [96]	Curated a dataset encompassing 450 images accompanied by text segmentation annotations, sourced from the Manga109 dataset.
2020	ECCV	DanbooRegion: an illustration region dataset [97]	Features 5377 pairs of region annotations, capturing diverse artistic region compositions.

(Continued)

Table 4 (continued)

Year	Conference /Journal	Title	Scale
2020	MMUL	Building a manga dataset "Manga109" with annotations for multimedia applications [98]	Comprises 109 Japanese manga volumes authored by 94 creators, totaling 21,142 pages, with over 500,000 annotations encompassing panels, text boxes, faces, and body poses. Upon its introduction, it represented the most extensive publicly available manga image dataset for research purposes.

2.4.2 Anime-Related Datasets

Table 4 provides an overview of key anime-related datasets. Despite advancements, current anime/manga datasets face notable limitations impeding sophisticated model development. Prominent among these is pervasive stylistic heterogeneity without granular annotation, hindering style-specific mastery or seamless generalization across diverse aesthetics [99,100]. This often results in dataset bias, over-representing popular styles and diminishing performance on less common ones, exacerbated by long-tailed character distributions.

Furthermore, the landscape is characterized by fragmentation, with datasets often task-specific and lacking comprehensive multi-modal integration of visual, textual, and structural elements crucial for holistic understanding [101]. Annotation quality and consistency remain concerns, prone to human error and complexity, particularly for intricate tasks [102]. Finally, the static nature of most datasets fails to capture the domain's dynamic evolution, limiting their enduring relevance for cutting-edge research. These constraints collectively underscore the pressing need for more nuanced, integrated, and continuously evolving data resources.

3 Image Generation

This section elucidates the current state of generative image synthesis models, focusing on prominent architectures particularly relevant to anime content creation, a domain where artificial intelligence is increasingly employed as an artistic tool. Our selection encompasses architectures distinguished by their efficacy in generating high-fidelity anime-style visuals, their impact on the field, and the depth of available research and practical applications. We focus on Stable Diffusion [3], related mainstream models, manga synthesis methodologies, and evaluation metrics.

While the substantial parameters, training compute, and VRAM requirements detailed in Tables 5 and 6 underscore the significant computational investment in these models, these costs are often offset by significant gains in creative efficiency and reduced manual labor. Moreover, ongoing research into architectural efficiencies and optimization techniques actively seeks to enhance their accessibility and integration into practical creative pipelines, continuously lowering the effective cost of deployment relative to the benefits derived.

Table 5: Mainstream image generation large models

Model Name	Rel. Date	Dev. Team	Params	Open Source
Stable diffusion 1.5 [3]	Oct 2022	StabilityAI	983M	Yes
Stable diffusion XL [3]	Jul 2023	StabilityAI	3.5B	Yes

(Continued)

Table 5 (continued)

Model Name	Rel. Date	Dev. Team	Params	Open Source
Stable diffusion 3.0 [53]	Feb 2024	StabilityAI	800M-8B	Yes
Stable diffusion 3.5 [53]	Oct 2024	StabilityAI	2.5B-8B	Yes
FLUX.1 [103]	Aug 2024	Black forest labs	12B	Yes
Midjourney V7 [104]	Apr 2025	Midjourney	Undisc.	No
DALL·E 3 [105]	Aug 2023	OpenAI	Undisc.	No
Imagen 3 [29]	Aug 2024	Google	Undisc.	No
Firefly [106]	Mar 2023	Adobe	Undisc.	No
Ideogram 3.0 [107]	Mar 2025	Ideogram AI	4B	No
CogView-4 [108]	Mar 2025	Zhipu AI	6B	Yes
Kolors [109]	May 2024	Kuaishou	3B	Yes
Seedream 3.0 [110]	Apr 2025	ByteDance	Undisc.	No
HunyuanDiT [111]	May 2024	Tencent	15B	Yes

Note: Estimated Training Compute:Stable Diffusion: 1.5~150,000 A100 h (based on v1-x estimation); Stable Diffusion XL: SD 2.0 estimated at 0.2M A100 h

Table 6: Resource efficiency and hardware demands of state-of-the-art image generative models

Model	Std. VRAM Req.	Opt. Min. VRAM	Optimization Techniques
Stable diffusion 1.5	4–6 GB	~2 GB	Attention slicing, low-VRAM Mode, Core ML
Stable diffusion XL	8 GB	~6 GB	FP8 Quantization, CPU offloading, Microsoft olive
Stable diffusion 3.0	12–24 GB (est.)	Not specified	FP16 quantization, Model offloading (assumed)
Stable diffusion 3.5	12–24 GB	~8 GB	FP8 quantization, CPU offloading, 8-bit DiT
FLUX.1	24 GB+	6 GB	FP8/NF4 quantization, System RAM offloading
HunyuanDiT V1.1	12–32 GB	~8 GB	Q4/Q5 GGUF, TensorRT, Flash-attention

3.1 Stable Diffusion Models

3.1.1 Image Synthesis via Stable Diffusion Models

Stable Diffusion, a prominent tool in anime image synthesis, leverages latent diffusion to generate high-fidelity images through iterative denoising in latent space. Developed by StabilityAI, CompVis, and Runway, its initial release in October 2022 marked a significant advancement. Subsequent iterations, including Stable Diffusion XL 1.0 (July 2023), Stable Diffusion 3.0 (February 2024), and Stable Diffusion 3.5 (October 2024), have progressively enhanced resolution, text alignment, and overall generative performance. Notably, the FLUX model (August 2024) [112], by Black Forest Labs, presents a competitive alternative, demonstrating superior image quality. Fig. 6 illustrates the evolution of image generation quality across Stable Diffusion versions.

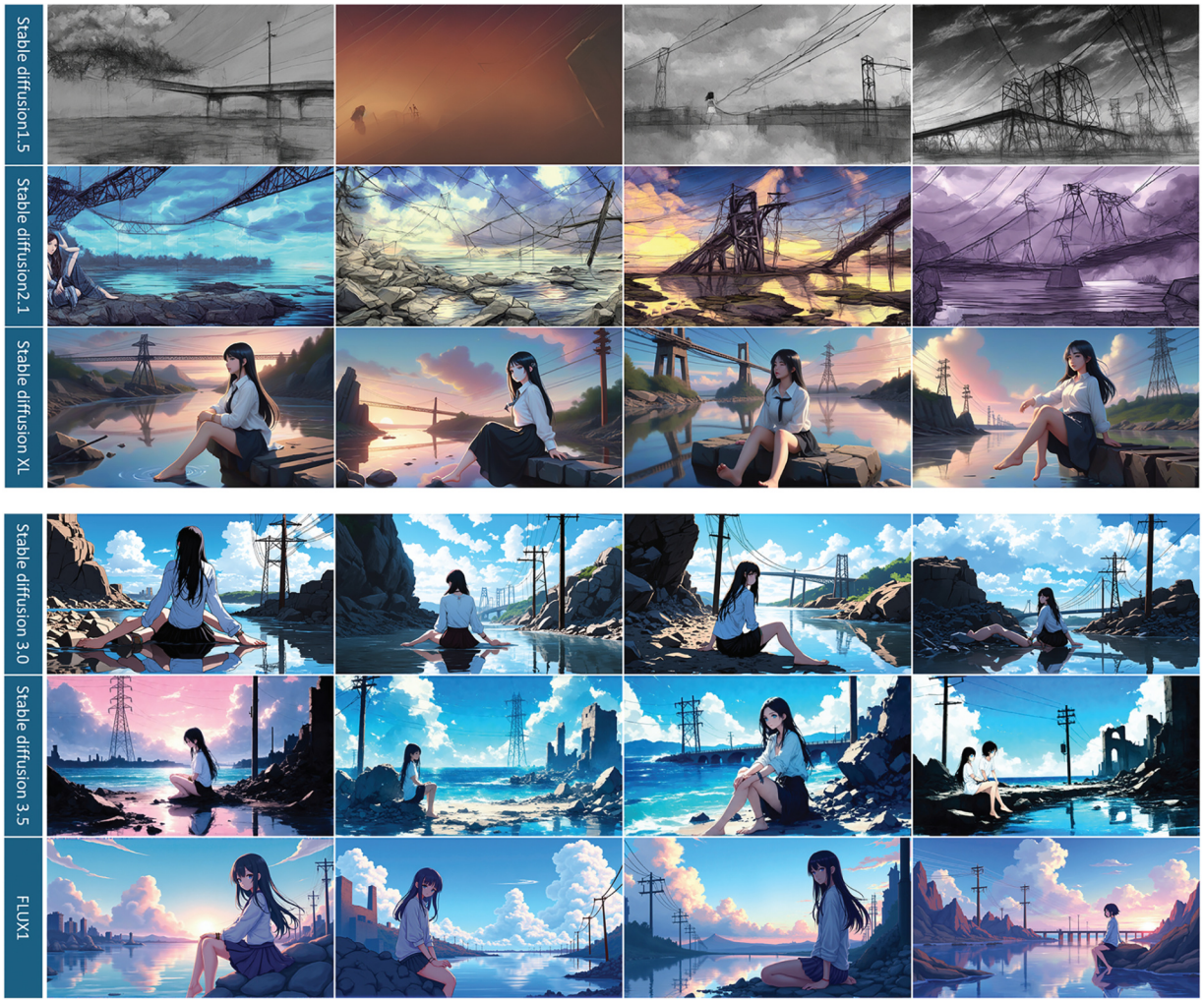


Figure 6: Evolution of Stable Diffusion Models demonstrated via images generated under consistent input conditions. See Appendix A

3.1.2 Ancillary Technologies in Stable Diffusion Ecosystems

The open-source nature of Stable Diffusion has fostered a diverse ecosystem of ancillary technologies, enhancing its applicability in artistic creation.

Controllable Image Synthesis: ControlNet (February 2023) enables precise pose and detail manipulation through trainable copies conditioned on edge and contour maps [78], As demonstrated in Fig. 7. T2I-adapter (February 2023) provides analogous functionality with a lightweight architecture [113]. MaskDiffusion (March 2024) refines textual controllability [114].

Model Fine-Tuning and Optimization: LoRA (June 2021) achieves efficient fine-tuning by reducing trainable parameters [73]. Latent Consistency Models (LCMs) (October 2023) accelerate fine-tuning by learning latent space mappings [115].

User Interface and Workflow Management: stable-diffusion-webui provides a user-friendly web interface for text-to-image synthesis [116]. ComfyUI offers a node-based interface for advanced customization and pipeline construction [117].



Figure 7: ControlNet controls the generation of images by adding constraints

Advanced Image Manipulation: PaintsUndo (August 2024) simulates painting brushstrokes and enables sketch extraction and anime-style transformations [118]. Its generation process is detailed in Fig. 8.



Figure 8: PaintsUndo model generation process

3.2 Other Models and Related Technological Landscape

3.2.1 Notable High-Impact Models

While these models collectively push the boundaries of generative capabilities, the landscape of text-to-image synthesis, beyond established open-source models like Stable Diffusion, is populated by a diverse array of other high-performing architectures.

ImagenFX (Google Labs), leveraging Imagen 2 and 3 [29,119], provides global users with high-fidelity visual content generation through an intuitive interface and robust semantic interpretation. Its integration with Gemini enhances multilingual support and complex scene rendering. Lumina-Image 2.0 (Alpha-VLLM) employs a 2.6B-parameter DiT architecture and a Gemma-2B text encoder [120],

demonstrating superior text-following performance in DPG and GenEval benchmarks [121–123], particularly for Sino-Anglic prompts. CogView3 (Zhipu AI) utilizes a cascaded diffusion framework and relay diffusion techniques [124], achieving accelerated inference and enhanced generative quality via a three-stage generation strategy and diffusion distillation optimization. Adobe Firefly [106], integrated within Adobe Creative Cloud and powered by Adobe Sensei, offers comprehensive image, video, and audio synthesis, trained on commercially compliant datasets, excelling in 1080p video generation. DALL·E 3 [62,125], through its integration with GPT, enhances intelligent image generation and editing, improving user interaction naturalness and editing precision. Ideogram focuses on generating images with legible textual elements, addressing a critical limitation in existing generative AI tools [107]. Its rapid synthesis and diverse stylistic options facilitate the creation of high-quality visuals for logo, poster, and graphic design.

Illustrious (OnomaAI Research) [126], built upon the SDXL architecture and leveraging Danbooru tags and multi-level captions, specializes in high-fidelity anime and illustration synthesis, excelling in resolution, color gamut, and anatomical accuracy. Pony Diffusion [127], an SDXL-derived anime model, has garnered significant acclaim on Civitai, recognized for its exceptional stylistic adaptation and high-resolution synthesis.

HunyuanDiT (Tencent) employs a DiT architecture with multi-resolution training and a dual-encoder system [52,111], demonstrating superior text-image consistency, subject clarity, and aesthetic fidelity, particularly for Chinese prompts.

These architectures, through continuous innovation, collectively propel the evolution of text-to-image synthesis. Regarding their specific relevance to anime generation, many proprietary models primarily leverage their robust general prior knowledge to infer and render anime aesthetics. In contrast, models such as Illustrious and Pony Diffusion exemplify targeted specialization, having been comprehensively fine-tuned on anime-specific datasets to significantly enhance the anime generation capabilities of their Stable Diffusion base, achieving remarkable stylistic adaptation and fidelity. It is also noteworthy that HunyuanDiT, an open-source model from Tencent built on the Diffusion Transformer (DiT) architecture, represents a convergence of language model and diffusion model strengths. However, as a comparatively recent entrant to the open-source domain, its subsequent development has often involved the adaptation and integration of techniques originating from the more established Stable Diffusion ecosystem. Fig. 9 compares images generated by various models under consistent conditions.

3.2.2 Related Technological Developments

Recent advancements significantly enhance **anime generation** by reducing costs, boosting efficiency, and refining control. **"Stretching Each Dollar"** democratizes high-quality generation through resource-frugal training [128], effective delayed patch masking, and synthetic data incorporation. While democratizing access, it struggles with precise text rendering and granular object control, particularly at high masking rates. **SANA** excels in high-resolution synthesis (up to 4K) via innovations like the deep compression autoencoder (AE-F32) and Linear DiT, improving prompt adherence with complex instructions through a decoder-only LLM [129]. Its speed and consumer hardware deployability lower entry barriers, but challenges remain in guaranteed content safety, controllability, and artifact handling in complex areas like faces and hands, showing a trade-off with raw reconstruction quality. Both accelerate rapid iteration in concept design and storyboarding; "Stretching Each Dollar" aids smaller studios, while SANA provides high-resolution assets and precise text-to-image alignment for diverse anime aesthetics.

ThinkDiff and **DREAM ENGINE** advance multimodal information fusion [130,131]. ThinkDiff integrates VLM outputs with diffusion processes for in-context reasoning, enabling complex instruction

interpretation and generation based on inferred relationships. Despite being lightweight and robust, it currently lacks high image fidelity and mastery of the full spectrum of complex reasoning tasks. DREAM ENGINE offers efficient text-image interleaved control through a versatile multimodal encoder and a two-stage training regimen, excelling in object-driven generation, complex composition, and free-form image editing. These capabilities are transformative for animation: ThinkDiff could automate storyboarding and character development, while DREAM ENGINE provides precise control for detailed scenes, character refinement, and visual consistency.



Figure 9: (Continued)



Figure 9: Depictions generated by various contemporary large-scale generative models using a consistent textual prompt and parameters. See Appendix A

For fine-grained control and targeted outputs, **MangaNinja** and **PhotoDoodle** are pivotal [132,133]. MangaNinja provides user-controllable diffusion-based manga line-art colorization, handling discrepancies between references and line art with a dual-branch structure and point-driven control. It enhances color consistency and detail, robust for complex scenarios, yet semantic ambiguity can arise with intricate line art, and a dependency on reference imagery persists. PhotoDoodle introduces an instruction-guided framework for learning artistic image editing from few-shot examples using an EditLoRA module. This enables efficient style capture from minimal data (30-50 examples) for seamless integration and mask-free instruction-based editing, though paired dataset collection and training are practical considerations. These tools significantly streamline labor-intensive processes: MangaNinja improves colorization efficiency and character consistency, while PhotoDoodle offers powerful stylistic application and precise, text-prompted modifications.

FluxSR optimizes image generation for practical applications like super-resolution via efficient single-step inference through Flow Trajectory Distillation (FTD) [134]. Built on powerful pre-trained text-to-image diffusion, it achieves superior perceptual quality and fidelity in recovering high-frequency details. Despite its high computational cost from a large parameter count and residual periodic artifacts, FluxSR offers transformative potential for enhancing low-resolution anime assets. This investment is justified by its ability to significantly streamline production through efficient upscaling of intermediate frames, ultimately contributing to a more rapid and less labor-intensive workflow.

Finally, **CSD-MT** reduces reliance on large labeled datasets through unsupervised content-style decoupling for facial content and makeup style manipulation [135]. Its efficiency, minimal parameters, and rapid inference offer flexible controls, yet extreme makeup styles can challenge accurate boundary rendering. CSD-MT's generalization to unseen anime makeup styles presents a pertinent avenue for efficiently designing and transferring makeup styles onto anime characters, streamlining visual development and creative exploration.

3.3 Sequential Image Synthesis for Manga Generation

Diffusion models [3], renowned for their efficacy in single-image synthesis, are demonstrating progress in generating coherent image sequences, which is a critical aspect of manga production.

Reference-Guided Synthesis: IP-Adapter (August 2023) guides diffusion processes using reference images, though with reduced textual prompt controllability [136].

Identity Preservation and Control: InstantID (January 2024) integrates facial and landmark images with textual prompts [137], imposing semantic and spatial constraints for identity consistency. PhotoMaker (December 2023) encodes multiple identity images into a unified embedding [138], preserving identity information while accommodating diverse identity integration. However, both models exhibit limitations in maintaining clothing and scene consistency across sequences.

Thematic Consistency Across Sequences: StoryDiffusion (May 2024) achieves thematic coherence within image batches by incorporating consistent self-attention into Stable Diffusion [139], facilitating manga-style narrative sequencing. The sequential narrative imagery synthesized by StoryDiffusion is shown in Fig. 10. Nevertheless, minor inconsistencies in character details persist in multi-character scenes.



Figure 10: Sequential narrative imagery synthesized via Storydiffusion

Multimodal Integration and Layout Control: DiffSensei (December 2024) integrates diffusion-based image generation with multimodal large language models (MLLMs) [140], employing masked cross-attention for seamless character feature incorporation and precise layout control. The MLLM-based adapter enables flexible character expression, pose, and action modifications aligned with textual prompts.

High-Fidelity Personalization: AnyStory (January 2025) utilizes an "encoding-routing" approach [141], employing ReferenceNet and an instance-aware subject router, to achieve high-fidelity personalization for single and multiple subjects in text-to-image generation. This approach enhances subject detail preservation and textual alignment.

While models like InstantID, PhotoMaker, and StoryDiffusion have shown promising advancements in generating consistent characters across scenes, minor inconsistencies in character details and clothing still persist in multi-character or long-form sequences, highlighting an area for continued research.

3.4 Evaluation Standards

Evaluation of generative image synthesis models bifurcates into automated and human-centric paradigms [142]. Automated metrics further diverge into content-invariant and content-variant assessments,

tailored to scenarios with and without ground truth, respectively [143]. Table 7 delineates the principal metrics and their salient characteristics.

Table 7: Evaluation metrics for image generation models

Metric category	Metric name	Description	Applicable scenarios
Content variation	Inception Score (IS) [144]	Evaluates quality and diversity using a pre-trained classifier (e.g., InceptionNet), measured by entropy. Lower entropy indicates higher quality.	Unconditional generative models, such as GANs
Content variation	Fréchet Inception DIstance (FID) [13]	Compares the distribution similarity between generated and real images in the feature space, based on the Fréchet distance.	Conditional and unconditional generative models
Content invariance	Learned Perceptual Image Patch SimIlarity (LPIPS) [145]	Computes perceptual similarity using VGGNet feature space, measuring structural changes.	Tasks with ground truth
Content invariance	Structural Similarity Index Measure (SSIM) [143]	Simulates human visual perception, measuring luminance, contrast, and structural similarity. Range from −1 to 1.	Tasks with ground truth
Content invariance	Peak Signal-to-Noise Ratio (PSNR)	Quantifies reconstruction quality through mean squared error (MSE). Higher values indicate better quality.	Tasks with ground truth

Beyond these foundational metrics, several nuanced evaluations merit consideration: Precision and Recall quantify the fidelity and completeness of generated samples; the F1 Score harmonizes these measures; Kernel Inception Distance (KID) [146] offers a robust alternative to FID using Maximum Mean Discrepancy; and the CLIP Score [14] evaluates semantic alignment between generated images and text prompts using pre-trained CLIP models, critical for text-to-image frameworks.

The discussion of generative models, particularly those fine-tuned or designed for specific aesthetics like anime, remains crucial. While empirical benchmarks are valuable, it's pertinent to acknowledge that standard metrics, such as the FID calculated with a general pre-trained Inception model, may not perfectly capture the perceptual nuances highly valued within specialized artistic styles. This underscores the im-portance of qualitative assessments and human-centric evaluations in understanding a model's true per-formance and aesthetic appeal, especially for complex and subjective domains like anime image synthesis. Further insights can be gained from analyzing representative sets of samples generated from diverse prompts, which can highlight models' varying strengths across multiple evaluation axes and demonstrate progress in producing perceptually appealing and contextually relevant imagery.

3.5 Summary of This section

The proliferation of generative models, notably Stable Diffusion and Flux, alongside community-driven plugins, has significantly expedited anime creation. Current creative paradigms primarily manifest in: (1) the rapid generation of evocative imagery within non-realistic domains, such as post-apocalyptic punk and steampunk, to aid creative ideation; (2) the meticulous refinement of model parameters to instantiate personalized stylistic generators, followed by iterative enhancement of initial outputs, encompassing aesthetic optimization, granular detail augmentation, structural anomaly rectification, and ambient modulation; and (3) the utilization of synthesized images as texture mapping repositories. However, if

the complexities of AI-generated content copyright are effectively navigated, and the synthesis process is judiciously controlled, iteratively refined, and finalized by human artists with keen aesthetic sensibility, the output of AI models can meet commercial production requirements. Furthermore, ancillary technologies facilitating the decomposition of image painting processes offer pedagogical utilities for education and novice practitioners.

4 Video Generation

The year 2024 marked a watershed moment in video synthesis, propelled by augmented computational resources and the refined capacity of generative models to process intricate spatiotemporal data. While still in a nascent stage compared to static image generation, this sub-field is rapidly transforming at the research and developmental frontiers. The advent of sophisticated open-source models has pushed the boundaries of fidelity, duration, and controllability. This section delves into prominent generative video synthesis models, reflecting the field's dynamic, albeit nascent, trajectory. Model selection emphasizes recent breakthroughs, novel architectural paradigms addressing core challenges like temporal coherence and computational efficiency, and demonstrable potential for application within anime production workflows. Notable exemplars, including Google's Gemini 2.0, OpenAI's DALL-E and Sora, Midjourney, and Meta's Make-A-Video, underscore the burgeoning technological sophistication within this domain [36,147].

4.1 Large Models for Video Generation

As Tables 8 and 9 illustrate, state-of-the-art video generation models involve even greater computational exigencies compared to their image counterparts, demanding substantial parameters, training resources, and VRAM. Addressing these considerable resource requirements is paramount for wider adoption, with ongoing research focusing on developing more efficient architectures and sophisticated optimization strategies to mitigate hardware demands and facilitate integration into complex animation workflows.

Table 8: Leading video generation large models

Model name	Params	Dev. team/Co.	Rel. date	Open source
Veo 3 [148]	Undisc.	Google deepmind	May 2025	No
Movie gen video [149]	30B	Meta AI	October 2024	No
Sora [147]	Undisc.	OpenAI	December 2024	No
LTX-Video [150]	2B/13B	Lightricks	May 2025	Yes
Mochi 1 [151]	10B	Genmo	October 2024	Yes
CogVideoX1.5 [152]	5B	Tsinghua Univ., Zhipu AI	November 2024	Yes
CogVideoX [152]	2B/5B	Tsinghua Univ., Zhipu AI	August 2024	Yes
HunyuanVideo [153]	13B	Tencent hunyuan	July 2024	Yes
Cosmos-1.0 [154]	7B/14B	NVIDIA	January 2025	Yes
SkyReels-V1 [155]	1.3B/13B	Kunlun world wide	February 2025	Yes
Wan2.1 [156]	1.3B/7B/13B	Alibaba	April 2025	Yes
Gen-4 [157]	Undisc.	RunwayML	April 2025	No
Pika 2.2 [158]	Undisc.	Glen pika	February 2025	No
Ray2 [159]	Undisc.	Luma labs	January 2025	No
Stable video diffusion 1.1 [160]	1.52	Stablility AI	February 2024	Yes
Kuaishou Kling 2.1 [161]	Undisc.	Kuaishou	May 2025	No
Hailuo 02 [162]	Undisc.	MiniMax	June 2025	No
Pyramid flow [163,164]	Undisc.	Kuaishou tech	October 2024	Yes

Note: Estimated Training Compute: Sora: Open-Sora 2.0 estimated ~$200 k training cost as a reference; Pyramid Flow: 20.7 k A100 GPU hours

Table 9: Resource efficiency and mitigation of VRAM demands in leading video generative models

Model	Std. VRAM Req.	Opt. Min. VRAM	Optimization strategies
LTX-Video	≥24 GB	8 GB+	FP8 quantization, multi-GPU parallelization
Mochi 1	≥24 GB	≥12 GB	FP8 scaled quantization, multi-attention backend
CogVideoX (2B)	18 GB	3.6 GB	INT8 Quantization, torch.compile
CogVideoX1.5 (5B)	26 GB	4.4 GB	INT8 Quantization, torch.compile
HunyuanVideo	≥32 GB	8 GB	Temporal tiling, LoRA
Cosmos-1.0	≥80 GB	≥39 GB (offloaded)	Model offloading strategies
SkyReels-V1	≥8 GB (1.3B)	≥8 GB	FP8 quantization, Multi-GPU parallelization
Wan2.1	≥8 GB (1.3B)	≥8 GB	Wan-VAE architecture
Stable video diffusion	≥80 GB	Not specified	Quality/Memory/Speed Trade-offs
Pyramid flow	12 GB	≥6 GB	CPU Offloading, Quantization

4.1.1 Capabilities and Challenges of Open-Source Generators

The advent of sophisticated open-source models has rapidly transformed video synthesis, pushing the boundaries of fidelity, duration, and controllability. A critical assessment of these advancements reveals diverse architectural strategies and performance profiles.

The **WAN model** leverages a dual image-video training paradigm to achieve broad synthesis capabilities, balancing efficiency and scale through its 1.3 and 14B variants. Its innovative spatiotemporal mechanisms facilitate the capture of complex dynamics, with the Streamer method enabling extended video generation at enhanced speeds. Nonetheless, challenges persist in preserving fine details during substantial motion and managing the computational cost associated with larger models.

Hunyuan Video, the largest open-source model evaluated at over 13 billion parameters, presents a comprehensive framework integrating advanced data curation and scalable training [153]. It excels in generating high-quality videos with precise text-video alignment and robust conceptual generalization. Its capabilities extend to coherent action sequences and localized text generation, partly attributed to large language model integration. While the original text doesn't explicitly detail disadvantages, challenges include the significant computational resources required and difficulties in maintaining perfect consistency in intricate scenarios. Artefacts from tiling during VAE inference require further refinement.

Stable Video Diffusion (SVD) marks a significant advancement in high-resolution latent video diffusion, underpinned by a systematic data management pipeline that elevates performance through pre-training [160]. SVD demonstrates a strong command of motion and 3D understanding, serving as a potent 3D prior for multi-view synthesis. Its modularity supports fine-tuning for downstream tasks like image-to-video generation and camera control. However, its efficacy is primarily limited to short videos, struggling with extended sequences, and inherent diffusion model characteristics result in slower sampling and high memory demands.

The **Pyramidal Flow Matching** approach enhances video generative modeling efficiency through a unified algorithm that reduces training tokens via temporal pyramids and integrates trajectories into a single Diffusion Transformer. This method supports high-quality video generation up to 10 s but can introduce subtle subject inconsistencies in longer videos. Current limitations include a lack of support for keyframe or video interpolation and an opportunity for improved fidelity to intricate prompts.

LTX-Video focuses on realtime video latent diffusion by optimizing a Transformer-based model with a high-compression Video-VAE [150]. This enables efficient spatiotemporal attention and high-resolution output directly in pixel space, achieving faster-than-realtime generation speeds. Yet, the high compression inherently limits fine detail representation, and performance can be sensitive to prompt clarity. The model currently focuses on short videos, with domain-specific adaptability remaining largely unexplored.

CogVideoX employs an expert Transformer within its diffusion model to generate continuous videos up to 10 s with strong text alignment and coherent actions [152]. It utilizes a 3D VAE for improved compression and fidelity, addressing the persistent challenge of flickering in generated video sequences. While scalable, aggressive compression can hinder convergence, and high-quality fine-tuning might slightly diminish semantic capabilities. Achieving long-term consistency with dynamic narratives is a persistent challenge. **SkyReels-A1** is specifically designed for expressive portrait animation using a video diffusion Transformer framework [155]. It excels at transferring expressions and movements while preserving identity, producing realistic animations adaptable to various proportions. The model handles subtle expressions effectively. Nonetheless, identity distortion, background instability, and unrealistic facial dynamics remain challenges, particularly with extreme pose variations.

Finally, the **Cosmos World Foundation Model Platform** provides a framework for building world models for physical AI, generating high-quality 3D consistent videos with accurate physical attributes through diffusion and autoregressive methods. Its open-source nature promotes accessibility. However, these models are in an early stage of development, exhibiting limitations as reliable physical simulators regarding object permanence, contact dynamics, and instruction following. Evaluating physical fidelity is also a significant challenge.

In summation, the open-source video generative landscape is marked by diverse, powerful models. Ongoing research is essential to address current limitations in fidelity, efficiency, and controllability, paving the way for broader applications and continued innovation.

4.1.2 Performance Evaluation and Anime Potential

Recent advancements in open-source video generative models have demonstrated notable superiority over established baselines and various contemporary models in rigorous paper evaluations. The WAN model, for instance, has consistently surpassed existing open-source and sophisticated commercial solutions across a spectrum of internal and external benchmarks, exhibiting a decisive performance advantage, notably outperforming models such as Sora, Hunyuan, and various CN-Top variants in weighted scores and human preference studies. Despite these impressive gains in fidelity and text alignment, challenges persist, particularly in achieving perfect consistency in long-form narratives, maintaining fine detail during significant motion, and fully meeting the high standards for stylistic precision and emotional depth inherent in complex anime productions. Similarly, Hunyuan Video has been shown to exceed the performance of prior state-of-the-art models, including Runway Gen-3 and Luma 1.6, alongside several prominent domestic models, particularly excelling in text alignment, motion quality, and visual fidelity assessments [153]. Stable Video Diffusion (SVD) has proven superior to models like GEN-2 and PikaLabs in image-to-video generation quality and significantly outperformed models such as CogVideo, Make-A-Video, and

Video LDM in zero-shot text-to-video generation metrics [160]. The PyramidFlow model has distinguished itself by surpassing all evaluated open-source video generation models on comprehensive benchmarks like VBench and EvalCrafter, achieving parity with commercial counterparts such as Kling and Gen-3 Alpha using exclusively public datasets. LTX-Video has demonstrated a considerable lead over models including Open-Sora Plan, CogVideoX (2B), and PyramidFlow in user preference studies for both text-to-video and image-to-video tasks [144]. CogVideoX-5B has shown dominance over a range of models including T2V-Turbo, AnimateDiff, and VideoCrafter-2.0 in automated evaluations and outperformed the closed-source Kling in human assessments. Furthermore, SkyReels-A1 has exhibited superior generative fidelity and motion accuracy compared to diffusion and non-diffusion models like Follow-Your-Emoji and LivePortrait, also achieving higher image quality than most existing methods [155]. Lastly, the Cosmos World Foundation Model Platform's components have showcased remarkable performance improvements, with its Tokenizer outperforming existing tokenizers like CogVideoX-Tokenizer and Omni-Tokenizer in key metrics, and its World Foundation Models demonstrating significant advantages over VideoLDM and CamCo in 3D consistency, view synthesis, camera control, and instruction-based video prediction [152]. These findings collectively underscore the rapid progress and increasing competitive edge of open-source initiatives in the video generation domain.

Capitalizing on their recent performance breakthroughs, these sophisticated video generative models constitute a significant development for anime production, and can augment the creative and technical palette available to animators. These models demonstrate capabilities ranging from generating diverse artistic styles and handling multi-language text integration to exhibiting robust generalization in avatar animation tasks, including anime and CGI characters, with precise control over pose and expression. The capacity for high-resolution, temporally consistent video generation from text or images provides valuable tools for preliminary concept visualization, storyboarding, and generating certain complex scenes or effects. This can augment workflow efficiency and reduce production overheads. Furthermore, specialized models excelling in expressive portrait animation with accurate facial and body motion transfer, adaptable to varied anatomies and scene contexts, directly address the nuanced demands of character animation in anime. While challenges persist in achieving perfect consistency in long-form narratives, maintaining fine detail during significant motion, and fully meeting the high standards for stylistic precision and emotional depth inherent in anime, the open availability and progressive capabilities of these models offer a fertile ground for developing next-generation animation techniques and tools. Their foundational strengths in high-quality video synthesis, 3D consistency, and increasing controllability suggest a potential for streamlining pipelines, fostering creative exploration, and contributing to the visual lexicon of anime.

Comparative video generation results from various large models, obtained on identical hardware using a uniform text prompt, are presented in Figs. 11–18.



Figure 11: CogVideoX 1.5 (5B Parameters). See Appendix A



Figure 12: Cosmos-1.0 (7B Parameters)



Figure 13: HunyuanVideo



Figure 14: LTX-Video (2B Parameters)



Figure 15: Mochi 1 (10B Parameters)



Figure 16: Pyramid-Flow



Figure 17: SkyReels-V1



Figure 18: WAN2.1

4.2 Contemporary Strides in Video Synthesis

Drawing upon recent breakthroughs in generative AI, particularly within diffusion and language models, the landscape of anime video synthesis is rapidly advancing. A critical challenge in this domain lies in effectively generating extended, multi-scene narratives with both temporal and character consistency, while simultaneously addressing the exponential increase in computational demands Contemporary research endeavors are addressing this complex challenge through innovative architectural designs and optimization strategies. Techniques such as global-local diffusion cascades and segmented cross-attention mechanisms are being developed to enhance efficiency in processing long video sequences. Alongside these architectural improvements, optimization strategies like temporal tiling and parameter-efficient fine-tuning (e.g., LoRA) are crucial for managing computational resources, particularly VRAM. Furthermore, significant strides are being made in explicitly improving temporal coherence through methods like consistent self-attention, temporal-aware positional encoding, and latent state variable-based modeling. Concurrently, advancements in character animation are focusing on maintaining consistent appearances across frames and scenes via appearance encoders, multi-scale feature fusion networks, and query injection mechanisms. While

challenges persist, the convergence of these architectural, algorithmic, and optimization-focused innovations is contributing to progress towards more efficient and consistent generation of long-form anime video content. A overview of these advancements is presented in Fig. 19.

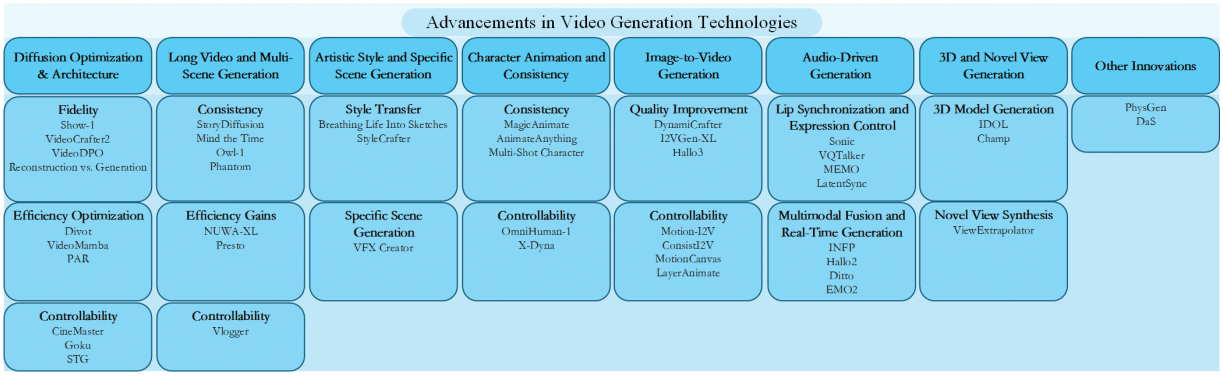


Figure 19: Overview of advancements in video generation technologies

Key advancements can be categorized into four interconnected areas, with particular emphasis on their implications for anime:

Diffusion Model Optimization and Architectural Innovations: Foundational to high-fidelity video generation, this area focuses on refining diffusion models for enhanced quality, efficiency, and controllability. Core contributions include hybrid pixel-latent diffusion frameworks (e.g., Show-1), high-fidelity image fine-tuning (e.g., VideoCrafter2), and human preference alignment [165–167]. For anime, these improvements directly translate to sharper visuals and more aesthetically pleasing outputs. Efficiency gains, such as parallel generation strategies (e.g., PAR), are crucial for scaling up anime production, which often involves numerous frames [39].

Long Video and Multi-Scene Generation: This domain tackles the critical challenge of maintaining temporal coherence and content consistency across extended narratives, a paramount concern for anime series and films. Pivotal methods include consistent self-attention and semantic motion predictors (e.g., StoryDiffusion), temporal-aware positional encoding, and latent state variable-based modeling (e.g., Owl-1) [139,168,169]. These innovations are directly applicable to ensuring narrative flow and visual continuity in multi-episode anime productions. Efficiency techniques like global-local diffusion cascades (e.g., NUWA-XL) and segmented cross-attention mechanisms (e.g., Presto) are vital for generating long anime sequences without prohibitive computational costs [170,171].

Character Animation and Consistency: Maintaining consistent character appearances and actions across frames and scenes is indispensable for compelling anime. Research here centers on achieving temporal coherence and visual realism. Seminal works utilize video diffusion models with appearance encoders (e.g., MagicAnimate), multi-scale feature fusion networks, and query injection for cross-shot consistency [172–174]. These advancements are directly responsible for preserving character identity and fluidity of motion throughout anime narratives.

Artistic Style and Specific Scene Generation: This area is particularly relevant to anime, which relies heavily on distinctive artistic styles. It focuses on achieving stylistic diversity and scene-specific content. Key methodologies involve T2V priors and deformation techniques and reference image-driven style adapters (e.g., StyleCrafter) [175,176]. These enable precise control over the aesthetic qualities of generated anime, allowing for the replication of diverse artistic styles and the creation of highly customized visual content.

4.2.1 Diffusion Model Optimization and Architectural Innovations

The adaptation of diffusion models to video generation, while promising, necessitates addressing inherent complexities related to temporal dynamics, computational efficiency, and user controllability. Current research endeavors are strategically focused on enhancing generation fidelity, optimizing computational throughput, and augmenting user-directed manipulation.

Regarding fidelity, seminal works have introduced hybrid pixel-latent diffusion frameworks (Show-1, October 2023) to balance quality and efficiency [166], refined spatial modules via high-fidelity image fine-tuning (VideoCrafter2, January 2024) [167], and leveraged human preference alignment through novel metrics (VideoDPO, December 2024) [165]. Furthermore, the reconciliation of reconstruction and generation objectives via VA-VAE and LightningDiT (Reconstruction vs. Generation, March 2025) *has shown* improvements in image quality and training efficiency [177].

For efficiency optimization, PAR (December 2024) introduces a parallel generation strategy by discerning dependencies among visual tokens [39], with the aim of augmenting the efficiency of image and video generation while maintaining the quality of autoregressive models.

Controllability has been enhanced through frameworks that enable cinema-level control over objects and cameras (CineMaster, February 2025) [178], rectified flow transformers achieving benchmark performance (Goku, February 2025), and transformer-based diffusion models that improve sample quality (STG, November 2024) [179].

These advancements collectively propel diffusion models towards a synergistic enhancement of quality, efficiency, and controllability, expanding technological frontiers through multimodal integration, spatial-temporal decoupling, and efficient architectural paradigms.

4.2.2 Long Video and Multi-Scene Generation

The generation of long-form and multi-scene videos poses significant challenges related to temporal coherence, content consistency, and computational scalability. Innovations in this domain emphasize consistency preservation, efficiency augmentation, and user-directed control.

Consistency is addressed via consistent self-attention and semantic motion predictors (StoryDiffusion, May 2024) [139], temporal-aware positional encoding (Mind the Time, December 2024) [169], latent state variable-based world evolution modeling (Owl-1, December 2024) [168], and thematic element extraction through cross-modal alignment (Phantom, February 2025) [180].

Efficiency gains are realized through global-local diffusion cascades (NUWA-XL, March 2023) and segmented cross-attention mechanisms (Presto, December 2024) [170,171], enabling the generation of extended video segments with reduced computational overhead.

Controllability is enhanced through LLM-driven multi-scene script generation pipelines (Vlogger, March 2024; VideoStudio, January 2024) [181,182], facilitating complex narrative construction and user-directed content creation.

Despite these advancements, challenges remain in maintaining long-term consistency, mitigating computational demands, and ensuring seamless multi-scene transitions.

4.2.3 Artistic Style and Specific Scene Generation

This domain focuses on achieving stylistic diversity and scene-specific content generation, catering to artistic creation and personalized customization. Research avenues encompass style transfer and generation, and specific scene synthesis.

Style transfer is facilitated by T2V priors and deformation techniques (Breathing Life Into Sketches, November 2023) and reference image-driven style adapters (StyleCrafter, September 2024) [175,176], decoupling content and style through pre-training and fine-tuning.

Specific scene generation is enabled by controllable diffusion transformers (VFX Creator, February 2025) [183], allowing for user-directed animated visual effects synthesis.

These methodologies expand the creative potential of video generation through innovative architectural designs and multimodal fusion.

4.2.4 Character Animation and Consistency

Character animation research centers on generating temporally coherent and visually realistic animations. Key objectives include consistency enhancement and controllability augmentation.

Consistency is addressed via video diffusion models and appearance encoders (MagicAnimate, November 2023) [173], multi-scale feature fusion networks and frequency domain stabilization (AnimateAnything, November 2024) [174], and query injection for cross-shot consistency (Multi-Shot Character, December 2024) [172].

Controllability is enhanced through diffusion Transformer-based frameworks (OmniHuman-1, February 2025) and zero-shot, diffusion-based pipelines with dynamic adapters (X-Dyna, January 2025), enabling realistic and contextually rich animations.

Future research aims to address consistency in complex motion and multi-character scenarios and optimize computational efficiency.

4.2.5 Image-to-Video Generation

I2V generation focuses on synthesizing dynamic videos from static images, addressing challenges related to motion inference and content consistency. Research is bifurcated into quality improvement and controllability enhancement.

Quality is improved through text-aligned image context projection and noise connection (DynamiCrafter, November 2023) [184], cascaded models with hierarchical encoders (I2VGen-XL, November 2023) [185], and identity reference networks (Hallo3, March 2025) [186].

Controllability is enhanced via motion field predictors and temporal attention (Motion-I2V, January 2024), spatio-temporal attention and noise initialization (ConsistI2V, July 2024), user-driven cinematic shot design (MotionCanvas, February 2025), and layer-specific control mechanisms (LayerAnimate).

These advancements propel I2V technology towards enhanced realism and user-directed control, though challenges remain in complex scene consistency and computational efficiency.

4.2.6 Audio-Driven Generation

Audio-driven generation synchronizes video content with audio, addressing challenges related to lip synchronization, facial expression control, and multimodal integration. Research is segmented into lip synchronization and expression control, and multimodal fusion and real-time generation.

Lip synchronization is improved via global audio perception and motion decoupled control (Sonic, November 2024), facial motion tokenization (VQTalker, December 2024) [187], memory-guided temporal and emotion-aware audio modules (MEMO, December 2024) [188], and audio-conditional latent diffusion models (LatentSync, December 2024) [189].

Multimodal fusion is enhanced through dual-aspect audio driving (INFP, December 2024) [190], improved patch deletion and noise enhancement (Hallo2, October 2024) [191], explicit motion space and streaming inference (Ditto, June 2024), and two-stage audio-driven virtual avatar generation (EMO2, January 2025) [192].

These technologies advance audio-driven generation towards enhanced realism, multimodal integration, and real-time applicability.

4.2.7 3D and Novel View Generation

This domain focuses on generating 3D models and novel views from 2D inputs, addressing challenges related to 3D structure inference and spatial consistency. Research encompasses 3D model generation and novel view synthesis.

3D model generation is facilitated by bi-modal U-Nets and motion consistency loss (IDOL, December 2024) and SMPL model-guided depth [193], normal, and semantic map fusion (Champ, June 2024) [194].

Novel view synthesis is enhanced by optimized SVD denoising (ViewExtrapolator, November 2024) [195].

These advancements drive the integration of 3D and novel view generation in creative applications.

4.2.8 Other Innovations

Innovations such as physical modeling integration (PhysGen, September 2024) and 3D tracking video-driven diffusion (DaS, January 2025) expand the boundaries of video generation by incorporating domain-specific knowledge [196,197].

4.3 Recent Innovations in Video Editing Methodologies

Fig. 20 provides an overview of recent advancements in video editing.

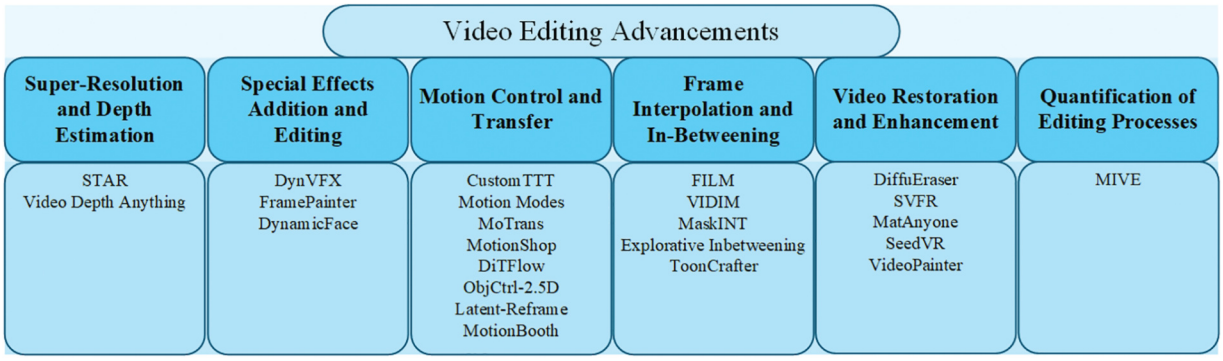


Figure 20: Overview of advancements in video editing technologies

4.3.1 Super-Resolution and Depth Estimation

Innovations in video editing are significantly propelled by advancements in super-resolution and depth estimation. STAR (January 2025) leverages a local information enhancement module and dynamic frequency loss [198], coupled with a text-to-video model, to achieve spatio-temporal enhancement in real-world video super-resolution, thereby improving detail fidelity and temporal coherence. Video Depth Anything (January 2025) introduces an efficient spatio-temporal processing head and concise temporal consistency loss [199],

enabling high-quality, temporally consistent depth estimation for ultra-long videos, achieving state-of-the-art zero-shot performance.

These methodologies advance video processing by enhancing visual fidelity and geometric perception. STAR balances detail and consistency through multimodal guidance and frequency domain constraints, while Video Depth Anything (January 2025) achieves zero-shot depth estimation with architectural efficiency [199]. Future research must address real-time performance and complex scene adaptability, critical for applications in film production and autonomous driving.

4.3.2 Special Effects Addition and Editing

The domain of special effects addition and editing has seen significant innovation. DynVFX (February 2025) employs a zero-shot, training-free framework [200], utilizing pre-trained text-to-video diffusion and vision-language models, to integrate dynamic content that interacts naturally with scenes based on textual instructions. FramePainter (January 2025) enables interactive image editing through intuitive visual interaction operations [201]. DynamicFace (January 2025) utilizes composable 3D facial priors, diffusion models [202], and temporal layers to achieve high-quality, consistent video face swapping, enhancing identity preservation and expression accuracy.

These advancements augment video/image editing capabilities by facilitating dynamic special effects generation, interactive editing, and identity consistency maintenance. DynVFX lowers the barrier to special effects production [200], FramePainter enhances creative flexibility [201], and DynamicFace resolves coherence challenges in face-swapped videos [202]. Future research must address controllability, complex scene adaptability, and ethical considerations, critical for film, advertising, and virtual content production.

4.3.3 Motion Control and Transfer

Motion control and transfer techniques are evolving to enable fine-grained dynamic editing. This shift facilitates the development of content generation tools with enhanced user control, physical plausibility, and usability. Research areas include dynamic parameter adjustment, motion-appearance decoupling, zero-shot motion transfer, 3D-aware motion control, and localized motion customization.

Dynamic parameter adjustment, exemplified by CustomTTT (December 2024) [203], utilizes test-time training to optimize appearance and motion LoRA parameters, mitigating artifact issues in multi-concept combinations. Motion-appearance decoupling, as seen in Motion Modes (November 2024) and MoTrans (December 2024) [204,205], independently models motion and appearance features, enabling decoupled control or transfer. Zero-shot motion transfer, demonstrated by MotionShop (December 2024) and DiTFlow (December 2024) [206,207], transfers motion patterns without target domain data. 3D-aware motion control, as in ObjCtrl-2.5D (December 2024) and Latent-Reframe (December 2024) [208,209], utilizes 3D geometric information to guide motion generation. Localized motion customization, as in MotionBooth (October 2024) [210], enables fine-grained editing of specific regions.

These technologies enhance dynamic controllability through parameter optimization, feature decoupling, zero-shot adaptation, and 3D perception. Future research must address complex interaction modeling, long-term temporal consistency, and ethical implications.

4.3.4 Frame Interpolation and In-Betweening

Frame interpolation and in-betweening are pivotal techniques for enhancing video smoothness and visual quality. Diffusion models, as in FILM (July 2022) and VIDIM (April 2024) [211,212], and Transformer architectures, as in MaskINT (April 2024) and EITS (March 2024) [213,214], have significantly advanced

this domain. ToonCrafter (May 2024) addresses non-linear motion and occlusion in anime videos through cartoon correction learning and a dual-reference 3D decoder [215].

The integration of diffusion models and Transformers has improved interpolation quality, particularly in complex scenes. Technologies like MaskINT indicate a trend towards real-time applications [213], while ToonCrafter exemplifies the deepening of technology in vertical domains [215].

4.3.5 Video Restoration and Enhancement

Video restoration and enhancement are essential for repairing damaged content and improving video quality. Diffusion models, as in DiffuEraser and SVFR (January 2025) [216,217], attention mechanisms, as in MatAnyone (January 2025) and SeedVR (February 2025) [218,219], and dual-stream architectures, as in VideoPainter (March 2025) [220], have advanced this domain.

These methodologies improve restoration quality and temporal consistency. The integration of multi-tasks, such as super-resolution, restoration, and colorization, and the development of optimization strategies, are key trends.

4.3.6 Quantification of Editing Processes

The MIVE (December 2024) framework quantifies editing leakage through the Cross-Instance Accuracy (CIA) score [221], decoupling Diverse Multi-instance Sampling (DMS) and Instance-center Probability Re-allocation (IPR), achieving state-of-the-art performance.

4.4 Advanced Paradigms in Video Comprehension

The inherent spatiotemporal complexity and high dimensionality of video data pose significant challenges beyond static image analysis, necessitating models capable of discerning both spatial configurations and intricate temporal dynamics across diverse tasks. While early approaches leveraged 3D Convolutional Neural Networks (3D CNNs) for processing short temporal segments [222], these architectures exhibited inherent limitations in capturing long-range dependencies and demonstrated computational inefficiencies [223]. Subsequently, the advent of Transformer architectures offered global contextual modeling capabilities [64]; however, their quadratic computational complexity with respect to sequence length constrained their applicability to high-resolution or extended-duration video sequences [224].

Recent advancements underscore a discernible trend toward unified, multimodal frameworks. Vitron (October 2024) introduced a pixel-level visual Large Language Model (LLM) capable of comprehensive visual intelligence [225], encompassing understanding, generation, segmentation, and editing across both image and video modalities, thereby addressing a broad spectrum of visual tasks from granular to abstract levels. Similarly, Sa2VA (February 2025) synergistically integrated SAM2 and LLaVA to achieve dense multimodal understanding of both static and dynamic visual content [226], facilitating tasks such as referring segmentation and multimodal dialogue. This trajectory signifies a paradigm shift from unimodal, task-specific architectures towards holistic frameworks that leverage shared representations for cross-modal, multi-task learning, thereby mitigating reliance on bespoke designs and enhancing overall performance.

In the pursuit of enhanced computational parsimony, VideoMamba (April 2025) presented a novel spatiotemporal modeling paradigm characterized by linear complexity [227], effectively circumventing the limitations inherent in both 3D CNNs and Transformer architectures, rendering it particularly efficacious for the analysis of protracted video sequences. Concurrently, architectural innovation is exemplified by Divot (December 2024) [228], a diffusion-based video tokenizer engineered to encapsulate both spatial and temporal feature hierarchies, supporting both video understanding and generative applications. These

diverse methodologies not only highlight the heterogeneity of contemporary architectural designs but also reflect an evolving research focus extending beyond mere comprehension towards sophisticated multimodal capabilities.

4.5 Evaluation Metrics for Generative Video Synthesis

The evaluation of generative AI video synthesis necessitates comprehensive metrics that quantify spatial quality (frame-level), temporal consistency (cross-frame), and semantic relevance (task adherence). The inherent temporal dimension of video synthesis introduces complexities beyond those encountered in image generation, mandating specialized evaluation paradigms. An overview of common evaluation metrics for video generation models is provided in Table 10.

Table 10: Evaluation metrics for video generation models

Metric category	Metric name	Description	Applicable scenarios
Spatial quality	Inception Score (IS) [144]	Evaluates single-frame quality and diversity using a pre-trained classifier, based on entropy calculation.	Unconditional video generation, frame-level evaluation
Spatial quality	Fréchet Video Distance (FVD) [229]	Compares the distribution similarity between generated and real videos using video feature extractors (e.g., I3D).	Conditional and unconditional video generation
Spatial quality	Structural Similarity Index (SSIM) [143]	Measures frame-level structural similarity, simulating human visual perception.	Tasks with ground truth
Spatial quality	Peak Signal-to-Noise Ratio (PSNR)	Quantifies frame-level reconstruction quality through mean squared error.	Tasks with ground truth
Spatial quality	Learned Perceptual Image Patch Similarity (LPIPS) [145]	Evaluates frame-level perceptual similarity using deep features, scalable to videos.	Tasks with ground truth
Temporal consistency	Frame difference	Computes pixel-level differences between adjacent frames, measuring temporal smoothness.	All video generation tasks
Temporal consistency	Optical flow consistency [230]	Evaluates inter-frame motion consistency using optical flow estimation.	Dynamic scene generation
Overall quality	Video Quality Metrics (VQM) [231]	Combines spatial and temporal factors to evaluate overall video quality.	Video streaming quality evaluation
Cross-modal consistency	CLIP score [232]	Evaluates semantic consistency between video content and text prompts using the CLIP model.	Text-to-video generation

Recent architectural advancements, notably the synergistic integration of autoregressive language models and diffusion models through Diffusion Transformer (DiT) architectures, exemplified by pioneering large models such as Sora, HunyuanDiT, and the high-performing WAN2.1, have significantly propelled the capabilities of text-to-video generation towards enhanced coherence and fidelity. These sophisticated models underscore the critical need for robust evaluation metrics capable of assessing their multifaceted

performance. Evaluation methodologies are broadly categorized into automated and human-centric metrics. Automated metrics encompass spatial, temporal, and cross-modal assessments.

Other Relevant Metrics:

Mean Opinion Score (MOS) [233]: A human-centric metric, employing a 1-to-5 scale to quantify subjective quality.

The evaluation of generative video models highlights the diverse strengths and limitations of current architectures. Metrics such as SSIM, PSNR, and LPIPS provide insights into spatial fidelity and perceptual similarity at the frame level, which are crucial for capturing intricate visual details. Temporal consistency, assessed via Optical Flow Consistency and Frame Difference, speaks to the models' ability to generate smooth and coherent motion sequences. Cross-modal alignment, indicated by the CLIP Score, underscores the semantic relevance of the generated content to input prompts. While no single model definitively outperforms others across all criteria, continuous advancements in these architectures are leading to significant progress in achieving high-quality, temporally consistent, and semantically accurate visual narratives.

4.6 Section Synthesis: Advancements in Video Generation

The year 2024 marked a period of significant advancement in video generation, driven by rapid increases in computational capacity, architectural innovations in generative models, and sophisticated modeling of complex spatio-temporal dynamics. This section has systematically delineated the salient technological advancements across ten pivotal research domains, ranging from model optimization and motion control to long-form video synthesis and artistic stylization, illustrating the comprehensive trajectory from theoretical conception to practical implementation.

Architectural and optimization strategies, particularly within diffusion models, have yielded significant enhancements in generation efficiency and fidelity through techniques such as pixel-latent space fusion, spatial-temporal module decoupling, and human preference alignment [234]. In the domain of motion control and customized generation, fine-grained manipulation of object motion and camera trajectories has been achieved through parameter decoupling, mixed score guidance, and zero-shot transfer methodologies [235]. Long-form and multi-scene video generation has emerged as a critical area of focus, with advancements in self-attention consistency, LLM-driven script generation, and latent state variable modeling addressing temporal logic and entity coherence challenges [139]. Furthermore, artistic style transfer via vector animation adaptation and style decoupling [236], image-to-video synthesis leveraging multi-stage motion prediction and noise optimization [237], frame interpolation utilizing bidirectional diffusion and masked Transformer models [213], character animation integrating optical flow guidance and frequency domain stabilization, audio-driven generation incorporating global perception and emotion modeling [238], 3D and novel view synthesis through joint video and depth map generation [237], and the exploration of multi-task unified models and physical simulation have collectively expanded the technological horizon [196].

Despite these advancements, video generation confronts persistent challenges, including the maintenance of global consistency in ultra-long videos, the accurate simulation of complex physical phenomena, the achievement of real-time generation efficiency in multimodal interactions, and the facilitation of personalized control in artistic creation. Future research directions, including model lightweighting, world model construction, and multimodal fusion, are poised to overcome these limitations, driving innovation in human-computer collaborative creation and immersive experiences.

5 Music Composition

This section delineates the emergent paradigm of AI-driven music composition, scrutinizing contemporary large-scale generative architectures, sophisticated synthesis methodologies, audio-visual transduction, and rigorous evaluation protocols. In contrast to the mature domain of image synthesis, diffusion-based music generation, though nascent, exhibits an accelerated developmental vector, indicative of its profound potential. This section delineates pivotal AI-driven music composition models, acknowledging this domain's emergent landscape and varied accessibility. Model selection prioritizes architectures exhibiting significant advancements in generating high-fidelity music and audio, focusing on capabilities pertinent to anime soundtracks and sound design, while considering both open-source availability and demonstrated influence or reported performance.

5.1 Leading Large Models for Music Generation

The landscape of music and audio generation has witnessed a transformative shift with the advent of large-scale generative models. As delineated in Table 11, several influential architectures have recently emerged, representing the current vanguard and significantly advancing the state of the art in computationally creative auditory content. While the domain of diffusion-based music generation is relatively nascent compared to image synthesis, its developmental trajectory is notably accelerated, indicating its potential. This section provides a critical overview of some of the most impactful large models in this rapidly evolving field, leveraging the provided information to evaluate their capabilities, limitations, and potential applications, particularly within contexts like anime production.

Table 11: Leading large models for music generation

Model name	Dev. team	Rel. date	Params	Open source
Music AI sandbox [239]	Google DeepMind	April 2025	N/A	No
Stable audio [240]	Stability AI	April 2024	13.2B	Yes
MuseNet [241]	OpenAI	April 2019	N/A	No
Suno AI v4.0 [242]	Suno AI	November 2024	N/A	No
V2A [243]	Google Deepmind	June 2024	N/A	No
MusicGen [244]	Meta	January 2024	1.5B	Yes
Music-01	Minimax	August 2024	N/A	No
ACE-Step [245]	ACE Studio, JieyueXingchen	May 2025	3.5B	Yes

Among the prominent models, Stable Audio Open, developed by Stability AI, warrants specific scrutiny. Released in April 2024 with 13.2 billion parameters and notably open-source, this model is primarily oriented towards text-to-audio generation, excelling in crafting high-quality stereo sound effects and field recordings at a professional 44.1 kHz sample rate. This capacity holds considerable promise for enriching the sound design and ambient audio layers in animated productions. Its accessibility, being runnable on consumer-grade GPUs, further democratizes its potential use in creative workflows. Furthermore, its support for generating variable audio lengths (up to 47 s) offers flexibility for diverse scene requirements. However, critical limitations exist: the model currently struggles with generating audio containing connected speech or understandable voice/singing, a significant impediment for anime narratives reliant on dialogue and vocal performances. Additionally, its performance in generating high-quality music is noted as limited compared to some state-of-the-art music-specific models, likely influenced by its training predominantly on Creative Commons licensed data. Empirical evaluation indicates its strength in sound generation (e.g., outperforming

several AudioLDM2 variants and AudioGen on AudioCaps FAD openl3) but shows it lags behind other Stable Audio iterations in instrumental music generation on datasets like Song Describer, though it performed slightly better than MusicGen on the latter.

Another pivotal contribution comes from Meta's MusicGen, a 1.5 billion parameter model released in January 2024 as an open-source project. MusicGen distinguishes itself as a single language model architecture capable of generating high-quality monophonic and stereophonic music conditioned on text descriptions or melodic inputs. This conditional capability offers enhanced control over the output. The model employs an efficient token interleaving pattern within a single-stage Transformer, eschewing the need for cascaded models. Extensive empirical evaluation, particularly comprehensive human assessments, positions MusicGen favorably against established baselines such as MusicLM, Mousai, and Riffusion, demonstrating superior output quality and adherence to textual prompts. The model's robust controllability via text and especially melody renders it particularly germane to anime production. The ability to generate scores that align with specific moods dictated by text or to conform to provided melodic structures offers granular control crucial for synchronizing music with on-screen action and narrative flow. Nonetheless, challenges persist, including limitations in achieving fine-grained control without heavy reliance on classifier-free guidance and potential biases stemming from the predominance of Western-style music in its training data.

Beyond these models for which detailed evaluations were provided, Table 11 lists several other notable large models contributing to the rapidly evolving field of music generation. These include Google DeepMind's Music AI Sandbox (May 2024) and V2A (June 2024), OpenAI's pioneering MuseNet (April 2019), Suno AI's Suno AI v4.0 (November 2024), and Minimax's Music-01 (August 2024). While detailed technical reports for many of these are not publicly available at the time of writing, their inclusion by prominent research entities underscores the escalating interest and investment in large-scale audio generation. Notably, ACE-Step from ACE Studio and JieyueXingchen, listed with 3.5 billion parameters and marked as open source with a prospective May 2025 release, represents a significant future development in this domain. However, it is important to note that, as of the current date, a formal technical report detailing its architecture and performance metrics is presently unavailable.

Collectively, these leading large models signify a paradigm shift in music and audio generation, moving towards more capable, versatile, and controllable systems. Their diverse strengths and ongoing development trajectories highlight the dynamic evolution of this field, with increasing potential for direct application in creative industries such as anime, enabling more sophisticated and tailored auditory experiences, despite current limitations in areas like nuanced vocal synthesis for some models.

5.2 Generative Synthesis Paradigms: Diffusion and Audio-Visual Interplay

5.2.1 Advancements in Music Generation

Fig. 21 provides a taxonomy of recent advancements in diffusion-based music generation. Diffusion models have significantly advanced music synthesis, achieving high-fidelity audio through progressive denoising and improving quality, controllability, diversity, and human-AI co-creation. They address limitations of traditional methods in balancing sonic fidelity and structural complexity. However, challenges remain in achieving human-level emotional depth, long-duration coherence, precise attribute manipulation, and reducing computational cost.

Recent innovations address these limitations. Stable Audio Open [240], Noise2Music [246], DiffRhythm [247], and TangoFlux exemplify the enhanced fidelity and efficiency attainable through diffusion and flow-matching techniques [248]. Multimodal conditioning, as demonstrated by Seed-Music [249], Music ControlNet [250], and JASCO [251], facilitates fine-grained control via text, symbols, and other modalities.

YuE and Both Ears Wide Open extend generative capabilities to lyrics-driven full-song synthesis and spatial audio [252,253], respectively. MusicMagus and SMITIN introduce sophisticated editing and intervention techniques [254,255], enabling nuanced manipulation of musical attributes.

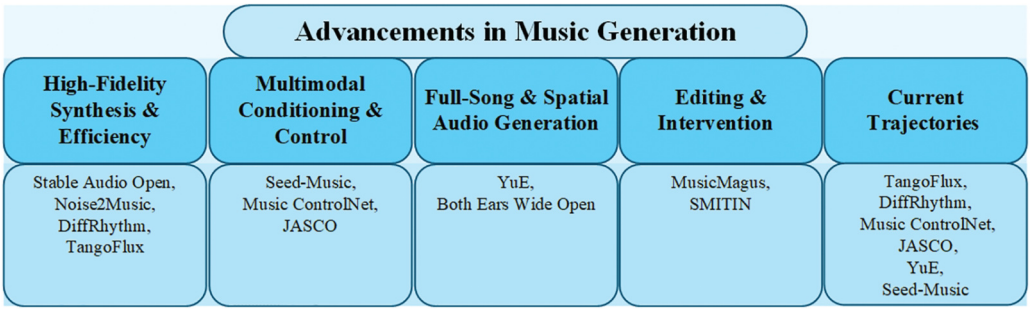


Figure 21: A taxonomy of recent advancements in diffusion-based music generation

Current trajectories underscore the pursuit of efficient, high-quality synthesis (TangoFlux, DiffRhythm) [247,248], personalized control (Music ControlNet, JASCO) [250,251], and multimodal integration (YuE, Seed-Music) [249,252].

5.2.2 Cross-Modal Audio-Visual Synthesis

Fig. 22 provides a taxonomy of recent advancements in cross-modal audio-visual music generation. Traditional music video (MV) production, characterized by intensive interdisciplinary collaboration, suffers from inherent inefficiencies, elevated costs, and challenges in cross-modal alignment. The escalating demand for personalized audio-visual content necessitates the development of efficient, automated, and artistically expressive cross-modal synthesis techniques.

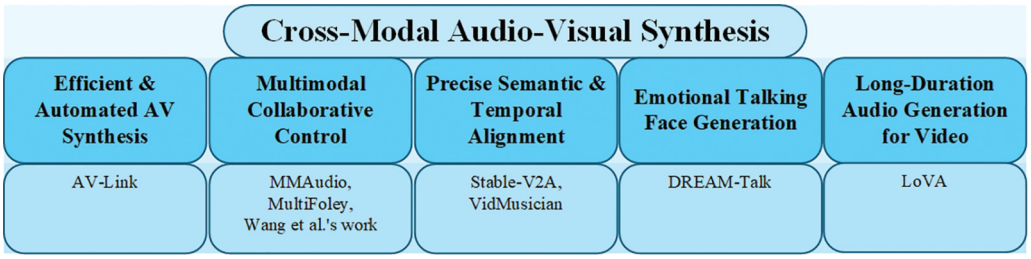


Figure 22: A taxonomy of recent advancements in cross-modal audio-visual music generation

Moving beyond unimodal generative paradigms, AV-Link introduces a unified framework leveraging temporally aligned diffusion features for bidirectional audio-video information exchange [256]. This architecture emulates human audio-visual cognition, facilitating dynamic modulation of inter-modal relationships.

Current research increasingly emphasizes multimodal collaborative control. MMAudio [257], MultiFoley [258], and Wang et al.'s work exemplify this trend [259], employing joint training and multimodal conditioning via video, text, and audio to enhance generative fidelity and cross-modal synchronization. This fusion amplifies creative latitude, enabling nuanced output manipulation through composite directives.

Global feature matching, however, often results in semantic coherence devoid of rhythmic concordance. Hierarchical modeling, as demonstrated by Stable-V2A and VidMusician [260,261], addresses this by

achieving precise semantic and temporal alignment. DREAM-Talk extends this precision to emotional talking face generation [262], utilizing diffusion models and video-to-video rendering for realistic expression and lip synchronization.

For extended video synthesis, LoVA employs a diffusion Transformer architecture to mitigate temporal inconsistencies in long-duration audio generation [263].

5.3 Datasets and Evaluation Metrics

The advancement of music generation and audio processing necessitates the development of robust datasets and evaluation paradigms. The FakeMusicCaps dataset addresses the critical issue of synthetic music attribution [264], facilitating audio forensics and mitigating copyright infringement. COCOLA introduces a consistency-oriented contrastive learning framework for music audio representation [265], enabling objective evaluation of harmonic and rhythmic coherence in generative models. The Sound Scene Synthesis Challenge, integrating objective and perceptual metrics, elucidates performance disparities across diverse sound categories and architectures. MixEval-X mitigates inconsistencies and biases in current evaluation methodologies by employing a multimodal benchmark and a mixed-adapt-correct pipeline [266], achieving modality-agnostic evaluation and enhancing alignment with real-world distributions.

Evaluation paradigms for generative music synthesis encompass spatial fidelity (pitch, rhythm), temporal coherence (rhythmic fluency), and semantic relevance (style, emotion) [267]. The inherent temporal complexity and artistic dimensionality of music necessitate specialized evaluation protocols, distinct from those employed in image or video synthesis [268]. A summary of these evaluation metrics is provided in Table 12.

Table 12: Evaluation metrics for music generation models

Metric category	Metric name	Description	Applicable scenarios
Subjective Eval.	Music auditory test	Evaluation through listeners' ability to distinguish AI-generated music from human-composed music, or through ratings based on criteria such as creativity and naturalness.	All music generation tasks
Subjective Eval.	Visual analysis	Expert analysis of waveform plots, spectrograms (e.g., Rainbowgram), musical scores, etc., to assess musicality.	Scenarios requiring expert analysis
Objective Eval.	Model-based metrics	Evaluation using metrics such as training loss, accuracy, F1 score, and chord prediction accuracy.	General generative model evaluation
Objective Eval.	Music Domain Metrics (MDM)	- Pitch and Rhythm: Scale consistency, pitch range, pitch count, etc. - Harmony: Harmonic consistency, chord entropy, etc. - Style: Style fitting, content preservation, etc.	Symbolic music (e.g., MIDI) evaluation
Objective Eval.	Fréchet Audio Distance (FAD)	Comparison of the similarity between the feature distributions of generated audio and real audio, based on the Fréchet distance.	Audio music generation evaluation
Comprehensive Eval.	Subjective + objective combination	Integration of listener ratings (e.g., harmony, rhythm) with objective metrics (e.g., polyphony, scale consistency).	Comprehensive evaluation
Comprehensive Eval.	Heuristic evaluation framework	Quantification of musicality using music theory tools (e.g., Circle of Fifths) and comparison with subjective testing results.	Symbolic music evaluation

Evaluation methodologies are broadly classified into automated (objective) and human-centric (subjective) assessments [268]. Automated metrics are further delineated into model-based, music domain-specific (MDM), and audio-specific metrics.

Beyond these core metrics, emerging and context-specific paradigms deserve consideration:

Structural Complexity Metrics: These automate the analysis of long-form musical structure by quantifying hierarchical segmentation, repetition rates, and variational patterns [269].

Semantic Content Metrics: Employing emotion/theme classifiers, these metrics assess the emotional and thematic fidelity of generated music, evaluating classifier performance through data augmentation [270,271].

Audio Quality Metrics: While less prevalent in music synthesis compared to speech, metrics such as spectrogram similarity and signal-to-noise ratio (SNR) provide crucial insights into sonic fidelity [272,273].

6 Game Generation

6.1 Basic Framework and Visual Effects of Games

6.1.1 Playable Game Generation

The advent of Artificial Intelligence Generated Content (AIGC) has catalyzed a paradigm shift in game development, advancing from unimodal image synthesis to complex multimodal video generation. However, the synthesis of interactive, high-fidelity playable games remains a formidable challenge. Current research trajectories encompass: diffusion-based methodologies (GameNGen) [274], integrating generative diffusion models with reinforcement learning for agent training; Transformer-based architectures (Oasis) [275], constructing open-world AI models for holistic game synthesis; DiT-based diffusion models (PlayGen) [276], achieving real-time interaction and accurate simulation of game mechanics; and large-scale foundational world models (Genie 2) [277], enabling the generation of infinitely diverse, controllable 3D environments for embodied agent training. These investigations underscore playable game synthesis as a pivotal AIGC direction, poised for further innovation. This section surveys nascent explorations into AI-driven game generation, concentrating on foundational advancements and integral components within this complex domain. Our selection highlights models and frameworks enabling high-dimensional content synthesis, such as 3D assets and environments, and those demonstrating the synergistic integration of diverse generative AI paradigms to address the inherent complexities of interactive game content creation. These sophisticated generative capabilities, enabling complex, high-dimensional content synthesis and multimodal integration for interactive experiences, demonstrate significant parallels and utility within diverse engineering and scientific fields [1]. Fig. 23 shows a real-time generated game screen from Oasis.



Figure 23: Oasis's real-time generated game screen

6.1.2 3D Scene Generation

3D scene synthesis constitutes a critical facet of game generation. Wonderland introduces a novel workflow for efficient high-fidelity 3D scene creation from monocular images [278]. DimensionX extends

this [279], generating realistic 3D and 4D scenes via video diffusion. Streetscapes demonstrates the generation of large-scale [280], consistent street-view imagery through autoregressive video diffusion, offering potential applications in urban planning and virtual reality. PaintScene4D generates consistent 4D scenes from textual prompts [281], advancing 4D synthesis. SynCamMaster enhances pre-trained text-to-video models for multi-camera video generation [282], ensuring inter-perspective consistency. Kiss3DGen (March 2025) repurposes 2D diffusion models for efficient 3D asset generation via "3D Bundle Images" [283]. Light-A-Video (February 2025) mitigates lighting inconsistencies in video relighting through a consistent lighting attention (CLA) module and progressive light fusion (PLF) strategy [284]. AuraFusion360 achieves high-fidelity 360° unbounded scene inpainting via depth-aware masking and adaptive guided depth diffusion [285]. Don't Splat your Gaussians (August 2024) introduces a novel volumetric modeling and rendering approach for scattering and emitting media [286]. Fast3R (January 2025) enhances 3D reconstruction efficiency via a Transformer-based architecture [287]. VideoLifter (January 2025) achieves efficient monocular video-to-3D reconstruction through a segment-based local-to-global strategy [288]. Gaga (December 2024) enables precise open-world 3D scene reconstruction and segmentation using 3D-aware memory banks. Bringing Objects to Life facilitates text-guided 3D object animation [289], preserving object identity. These advancements provide robust technical foundations for game development.

6.1.3 3D Character Synthesis

3D character synthesis represents a significant vector in game generation, focusing on the creation of realistic, animatable 3D character models. AniGS introduces animatable Gaussian avatars from monocular images [290], enabling real-time 3D puppet animation. Consistent Human Image and Video Generation employs spatial conditional diffusion for appearance consistency. Pippo (February 2025) achieves multi-view generation from monocular images via multi-stage training and attention bias [291]. PERSE (December 2024) synthesizes animatable personalized 3D avatars from single portraits [292], supporting facial attribute editing via decoupled latent spaces. These innovations bolster game character generation and animation production.

6.2 Specialized Game Mechanics via Generative AI

6.2.1 Autonomous Character Behavior

Autonomous character behavior constitutes a critical research vector in game AI, focusing on the synthesis of realistic and intelligent virtual agents. Current non-player characters (NPCs) often exhibit limitations in seamless game integration. Berkowitz (the director of Curiouser Institute) identified a central challenge: insufficient controllability of large language model (LLM)-driven AI, rendering their behaviors unpredictable and misaligned with game design specifications.

To address these constraints, researchers have explored several avenues. Motion Tracks introduces a 2D trajectory-based action representation and Motion Track Policy [293], enabling imitation learning from human video data for robotic agents. Generative Agents conceptualizes computational agents capable of simulating credible human behaviors [294], exhibiting both individual agency and emergent social interactions within simulated environments. Google DeepMind's SIMA interprets natural language directives and executes tasks within diverse 3D video game contexts [295].

6.2.2 Dynamic Character Customization

Dynamic character customization, specifically outfit alteration, is a salient research area in game AI, enhancing player immersion through personalized avatar appearance. Dynamic Try-On employs dynamic

attention mechanisms for virtual video try-on [296], preserving garment detail and temporal consistency during complex motions. This methodology offers a robust framework for realistic character outfit customization in gaming applications.

6.2.3 Immersive Driving Simulation

Immersive driving simulation represents a significant research domain within game AI, aiming to provide authentic and engaging vehicular experiences. The Stag-1 model facilitates the reconstruction of real-world driving scenarios and the synthesis of controllable 4D driving simulations [297], offering a novel approach to autonomous driving simulation. This methodology enhances the realism and interactivity of driving simulations, presenting potential for widespread adoption in gaming environments.

6.3 Industrial Deployment of Generative AI in Game Development

Despite the rapid proliferation of diffusion models and large language models in research and open-source communities, the comprehensive industrial deployment of generative AI within sectors like gaming and mainstream anime production remains largely nascent. Current implementations are primarily confined to localized artistic workflows or experimental projects. Beyond diffusion-based methodologies, alternative generative techniques are emerging for game asset creation. For instance, Tencent AI Lab's GiiNEX [298], showcased at GDC 2024, exemplifies this trend, achieving a substantial reduction in urban modeling time from five days to 25 min. Microsoft Research's "Muse" project pioneers the generation of playable gameplay sequences by learning intricate game dynamics and player interactions from extensive gameplay data [299]. Layer AI is refining mobile game design workflows by leveraging durable AI asset generation pipelines that ensure the reliability and consistency of content creation [300]. Scenario AI empowers game developers with an AI-driven platform to generate a plethora of game assets with consistent stylistic control [301], thereby accelerating the art production lifecycle.

Table 13: Leading large language models

Model name	Dev. team	Rel. date	Params	Category
Gemini 2.5 Pro [302]	Google	March 2025	N/A	MultiModal
Claude 3.7 [303]	Anthropic	February 2025	N/A	Multimodal
DeepSeek-R1 [304]	SenseTime Res.	January 2025	N/A	Multimodal
Mistral Large [305]	Mistral AI	July 2024	N/A	Multimodal
GPT-4.5 [306]	OpenAI	February 2025	N/A	Multimodal
Gork v4.0 [307]	xAI	July 2025	N/A	Multimodal
Pixtral Large [308]	MistralAI	November 2024	12B	Multimodal
Llama 3.2 [309]	Meta	July 2024	1B/3B/11B/90B	LLM
Mixtral-of-Expert [310]	MistralAI	April 2024	22B	LLM
Chatglm 4.0plus [311]	Zhipu AI	August 2024	N/A	LLM
Gemma 3.0 [312]	GoogleDeepmind	March 2025	1B/4B/12B/27B	LLM
Phi v4.0 [313]	Microsoft	December 2024	14B	LLM
DBRX [314]	Mosaic AI	March 2024	132B Tot, 36B Act	LLM
Nemotron-4 [315]	Nvidia	June 2024	340B	LLM
Hunyuan [316]	Tencent	February 2025	N/A	LLM
Hunyuan-T1 [317]	Tencent	November 2024	389B Tot, 52B Act	Multimodal
Doubao [318]	ByteDance	N/A	N/A	LLM

(Continued)

Table 13 (continued)

Model name	Dev. team	Rel. date	Params	Category
Kimi K2 [319]	Moonshot AI	N/A	N/A	LLM
Tongyi Qianwen [320]	Alibaba cloud	N/A	N/A	LLM
Qwen2 [321]	Alibaba cloud	Sep 2024	0.5B/1.5B/7B/57B	LLM
MiniMax [322]	MiniMax	January 2025	456B Tot, 45.9B Act	LLM

7 Alternative Applications

7.1 Generative Narrative Synthesis

Table 13 presents a list of leading large language models. Despite the rapid evolution and increasing scale of Large Language Models (LLMs), the automated synthesis of coherent and sustained long-form narratives continues to be a significant challenge. While these models exhibit evolving capabilities in generating fluent prose and localized textual segments and demonstrate progress in text coherence, fundamental limitations persist in managing extensive contextual dependencies, maintaining global narrative consistency, executing complex plot trajectories, and authentically representing nuanced affective arcs. These limitations impede the generation of truly cohesive and compelling stories suitable for complex anime narratives. These inherent constraints often lead to semantic drift, where the narrative coherence degrades over extended passages, leading to inconsistencies in plot, character, and theme [323].

Consequently, while LLMs show potential as tools within the creative writing pipeline, facilitating aspects such as genre-specific content generation or initial drafting, their effective deployment in producing complex narratives currently requires a robust, manually augmented, multi-stage process. This workflow typically includes iterative phases of human-led conceptualization, detailed structural outlining, and meticulous post-generation editing and refinement to compensate for the models' deficiencies in maintaining sustained narrative integrity.

Contemporary LLMs exhibit varying degrees of proficiency in tackling the complexities of narrative generation. For instance, Gemini 2.0 demonstrates enhanced logical coherence and temporal continuity in generated text compared to previous iterations [324,325], yet further advancements are *still needed* for reliably producing lengthy, intricate narratives without human intervention. Claude 3.7, while often lauded for its capabilities in character development and adherence to explicit instructions, can yield formulaic content and struggles with the intricate demands of sophisticated plot construction over long durations. DeepSeek-R1 showcases commendable prose quality and creative output [304], but is similarly susceptible to logical inconsistencies and deviations from initial narrative directives. Claude 3.5 Sonnet offers robust instruction following and maintains reasonable logical consistency [326], though perhaps at the expense of pronounced creative flair. Even prevalent models like GPT 4.0 [66], while versatile for a range of textual tasks, are often less suited to the demands of novel-length writing due to a tendency towards formulaic structures and only moderate long-range coherence, making them more suitable for shorter-form or professional content generation.

These observations underscore that while notable progress has occurred, particularly in localized text generation and adherence to immediate prompts, the significant challenge of synthesizing extensive, logically consistent, emotionally resonant, and structurally sound narratives without substantial human oversight largely remains unaddressed. Overcoming the persistent issues of semantic drift and maintaining global coherence are crucial challenges for future research in generative narrative synthesis.

7.2 Virtual Streamers

The evolution of virtual avatars in interactive media has progressed from software-synthesized vocaloids to human-operated VTubers. A new phase is emerging with autonomous AI-driven virtual streamers like Neuro-sama [327]. These entities utilize advanced AI, including deep learning and natural language processing, for real-time interaction, gameplay, and dynamic responses to viewers. Neuro-sama, the first female AI VTuber, debuted on Twitch in December 2022, showcasing fluent conversational engagement powered by large language models. This paradigm shift indicates a trend towards more personalized and intelligent virtual streaming experiences.

Luna AI exemplifies a state-of-the-art autonomous virtual avatar platform, integrating a suite of high-performance AI models, including GPT, Zhipu AI et al. These models facilitate both local and cloud-based deployments [328].

8 The Ethical and Progressive Arc of Generative Intelligence

8.1 Ethical and Security Implications of Generative AI

Generative Artificial Intelligence (Gen-AI) is significantly influencing user safety, digital content creation, and online platform governance. A salient trend is the accelerating transition of Gen-AI from a theoretical construct to a more pervasive technology in research and specialized applications, presenting both transformative opportunities and critical challenges for its comprehensive industrial adoption [329].

Gen-AI's capacity for nuanced natural language understanding renders it a pivotal tool in user safety. Research indicates its efficacy in detecting sophisticated violations, such as phishing and malware, which are challenging for traditional machine learning paradigms [329]. However, the transformative potential of Gen-AI, exemplified by models like GPT-4o and DALL-E 3, requires rigorous ethical scrutiny. While enabling high-fidelity content creation, Gen-AI introduces concerns regarding bias, authenticity, and potential misuse [330]. Online platforms, grappling with voluminous content, require advanced content moderation mechanisms. Researchers are developing collaborative frameworks leveraging annotation disagreement to address the subjectivity and complexity of toxicity detection [331]. Concurrently, they are exploring the legal, ethical, and practical challenges of AI-driven content moderation, emphasizing algorithmic bias, transparency, and accountability [332]. The increasing verisimilitude of AI-generated content necessitates robust detection methodologies. Interviews with Reddit moderators highlight the threats posed by AI-generated content and the limitations of current detection tools [333]. Research is also investigating techniques and challenges in AI-generated text detection, proposing future research directions [334]. The utilization of synthetic data in model training necessitates vigilance against potential biases. Research underscores the importance of synthetic data while acknowledging its inherent challenges and ethical implications [335]. Growing societal concern regarding the regulation of Gen-AI technologies underscores the need for robust governance frameworks. Research suggests parallels between Gen-AI and social media regulation [336]. The global impact of AI necessitates cross-cultural evaluation. Research highlights geographic biases in current studies and calls for broader global perspectives [337]. Population-specific impacts of AI require targeted research and intervention. Studies address the psychological impacts of algorithmic social media on adolescents, advocating for protective measures [338]. Ethical concerns regarding AI-generated text in scientific research, including transparency, bias, and accountability, are also being addressed [339,340]. The scalability of content moderation and the efficacy of AI-driven solutions are subjects of ongoing debate [341].

8.2 Specific Ethical Considerations in Anime and Cultural Content Production with Generative AI

Beyond the broad ethical implications of Gen-AI, its application within the specialized domain of anime and cultural content creation introduces a distinct set of considerations that warrant meticulous examination. The integration of generative AI into creative workflows, while augmenting production efficiency and fostering some expanded avenues for artistic expression, concurrently presents complex ethical dilemmas concerning authorship, cultural sensitivity, copyright, and the potential impact on traditional artistic communities [342]. Conversely, Chinese jurisprudence, under the Interim Measures, underscores human authorship, yet recent rulings permit copyright for AI-generated content with substantial human creative input, indicating a nuanced approach to co-creation.

Firstly, authorship becomes increasingly ambiguous when AI systems generate or co-create anime content. The traditional understanding of a singular human creator is challenged, raising questions about attribution, creative intent, and responsibility [343,344]. Current policy frameworks are beginning to address this. For instance, the EU AI Act, with full applicability by mid-2026, mandates transparency for generative AI, requiring disclosure that content is AI-generated and labeling of specific files [345]. Similarly, the UK's consultation proposes transparency from AI firms regarding training data and AI-generated content labeling [346]. However, these frameworks do not fully resolve the philosophical question of creative intent when AI co-creates, leaving a significant gap in defining ultimate accountability. China's Interim Measures (Article 4) mandate adherence to societal morality and socialist values, requiring AI-generated content to avoid cultural insensitivity and objectionable material (Article 10), albeit with broad guidelines for anime.

Secondly, the pervasive nature of anime as a global cultural phenomenon necessitates acute attention to cultural sensitivity. Generative AI models, trained on vast datasets that may reflect inherent biases, risk perpetuating stereotypes, misrepresenting cultural nuances, or inadvertently generating inappropriate content [347,348]. While existing legal frameworks, such as the EU AI Act, do not explicitly address cultural sensitivity, they include transparency requirements for training data summaries, which could indirectly aid in identifying and mitigating biases [345]. The UK's proposals for transparency in training data could also support cultural sensitivity by ensuring AI developers disclose datasets [346]. Conversely, Japan's developer-friendly AI rules, which permit training on copyrighted works without infringement, may exacerbate cultural sensitivity issues if biased data is utilized, as artists face significant challenges in proving misuse [349]. China's Interim Measures (Article 4) mandate adherence to societal morality and socialist values, requiring AI-generated content to avoid cultural insensitivity and objectionable material (Article 10), albeit with broad guidelines for anime.

Thirdly, the intricate landscape of copyright law faces significant challenges with the proliferation of AI-generated anime [350]. Questions arise regarding the copyrightability of AI-generated works, especially if the AI is trained on copyrighted material without explicit permission [351]. Global approaches to this issue diverge considerably. The EU AI Act requires generative AI to comply with EU copyright law and mandates the publication of summaries of copyrighted data used for training [345]. The UK is currently consulting on a copyright exception for AI training on copyrighted materials for commercial purposes, allowing rights holders to reserve rights [346]. In contrast, US copyright law typically requires human authorship, precluding the copyrightability of purely AI-generated works [352]. Japan's 2018 copyright law revision permits AI training on copyrighted works as "information analysis" without infringement, a stance that largely benefits developers but has sparked controversy among artists and legal experts due to potential implications for creator rights and proof challenges in court [349]. In China, courts recognize copyright for AI-generated works with human creative input, and the Interim Measures (Article 9) compel AI service providers to comply with IP law and ensure lawful training data.

Finally, the burgeoning capabilities of generative AI pose a significant potential impact on traditional artistic communities within the anime industry [353]. While AI can democratize content creation, there is a legitimate concern that it could lead to job displacement for traditional animators, illustrators, and voice actors [354–356]. Policies are evolving to mitigate this. The EU AI Act's emphasis on transparency and copyright compliance aims to protect creators by ensuring fair use of their work. Similarly, the UK's proposals underscore creator control and payment, supporting licensing deals to ensure artists can derive revenue from AI utilization [345,346]. However, Japan's more developer-friendly rules may exacerbate job displacement, as evidenced by instances such as Netflix's use of AI-generated background art, which generated global criticism for displacing animators [349,357]. The balance between fostering innovation and safeguarding artist livelihoods remains a contentious issue in policy development, underscoring the necessity for nuanced regulations. While China's Interim Measures (Article 5) foster AI innovation in cultural production, they do not explicitly address job displacement; however, the judicial emphasis on human authorship offers some indirect safeguard for traditional artists, reflecting a balance between innovation and IP protection.

8.3 Review of Progress and Foundational Challenges

Significant progress has been achieved across various domains of generative artificial intelligence. In image synthesis, diffusion models, exemplified by Stable Diffusion and its successors (e.g., SDXL 1.0, Stable Diffusion 3.0, FLUX) [3,38,67], have reached a degree of maturity, demonstrating considerable utility in artistic workflows through enhanced resolution and controllability via mechanisms like ControlNet and LoRA [73,250]. Nevertheless, a persistent technical challenge lies in maintaining rigorous inter-frame consistency in sequential image generation, particularly within complex, multi-character scenes. Standard evaluation paradigms, including Inception Score and Fréchet Inception Distance [13,144], while informative, often inadequately capture these temporal coherence issues.

The year 2024 marked a significant period for video synthesis, with the emergence of large-scale architectures such as Veo [269], Sora [147], and Movie Gen Video [149]. These models have substantially advanced generative quality and parametric control. Optimization strategies, including pixel-latent space fusion and spatio-temporal module decoupling [166,167], along with architectural innovations aimed at computational efficiency [227,228], have propelled the field forward. Crucially, however, despite advancements in fidelity and motion quality, achieving robust long-duration video consistency and real-time performance remains a significant bottleneck, demanding further exploration of advanced architectural paradigms. These limitations underscore that while the technology shows transformative potential, practical integration into fast-paced production pipelines requires overcoming these specific technical hurdles.

In music composition, diffusion-based models like Stable Audio Open and Noise2Music have elevated audio fidelity and control [240,246]. Yet, considerable hurdles persist in generating extended compositions with profound emotional depth and intricate structural coherence. Furthermore, cross-modal audio-visual transduction, while showing promise in enhancing synchronization, often incurs substantial computational overhead [256,261].

Game generation has witnessed notable strides in synthesizing playable content [274,276], 3D environments, and characters [278,279,290,291]. Despite this, developing interactive AI character behaviors with genuine autonomy and seamless integration remains constrained by limitations in generative fidelity and consistency [293,295].

Narrative synthesis, primarily driven by LLMs [324,326], continues to grapple with generating long-form content that transcends formulaic structures and avoids semantic drift or discontinuous plots. While autonomous virtual avatars highlight the potential for AI-driven real-time interaction [327], the underlying narrative generation capabilities require further maturation.

8.4 Pragmatic Hurdles in Deploying Generative AI for Anime Production

The assimilation of generative AI into the broader anime production sphere, while theoretically transformative, encounters significant practical challenges that temper its immediate widespread adoption, including considerable economic outlays for high-performance computing infrastructure, the need for specialized technical skills, and the complexities of integrating these tools into existing production pipelines.

Foremost among these are the considerable economic outlays. The initial procurement or development of sophisticated AI models, coupled with the requisite high-performance computing (HPC) infrastructure—encompassing powerful GPUs/TPUs and scalable cloud resources—represents a significant capital investment. Furthermore, ongoing operational expenditures, including energy consumption, data storage, model maintenance, and software licensing, contribute to a substantial total cost of ownership. However, these significant costs must be weighed against the potential for transformative long-term gains in production velocity, the reduction of highly specialized manual labor, and the unprecedented scalability of content generation that these technologies offer. For larger studios, this represents a strategic investment, while for smaller entities, continued research into more accessible and optimized solutions is crucial to unlock similar benefits [358,359].

Secondly, the infrastructural prerequisites extend beyond mere computational power. Robust data management pipelines are indispensable for curating, processing, and versioning the vast datasets essential for training and fine-tuning bespoke anime-centric AI models. High-bandwidth networking is also critical for efficient data transfer and collaborative workflows. The absence of, or deficiencies in, such infrastructure can significantly limit the utility and scalability of generative AI solutions [360].

Finally, a critical limitation resides in the specialized human capital required. Effective deployment necessitates a multidisciplinary workforce adept in AI/ML, data science, and software engineering, alongside artists and directors skilled in navigating and creatively leveraging these novel tools. A significant skills gap exists between traditional animation competencies and the emerging demands of AI-driven production, necessitating substantial investment in training, upskilling, or the recruitment of specialized talent. The seamless integration of AI into entrenched production pipelines also presents a significant workflow re-engineering challenge, requiring a notable shift in both creative and technical processes [361,362].

Addressing these financial, infrastructural, and skill-based constraints is pivotal for unlocking the full potential of generative AI within the anime industry and transitioning it from a nascent technology to an integral component of mainstream production.

8.5 Practical Integration and Adoption across Creative Industries

While theoretical advancements in generative AI models for anime production are significant, a critical evaluation of their practical adoption by studios and individual creators reveals a growing trajectory of integration, yielding tangible benefits and novel creative paradigms. The initial computational investments, while substantial, are increasingly offset by enhanced production efficiency, novel artistic expression, and expanded accessibility.

Evidencing this trend, diverse applications span across various creative domains. In visual arts, the award-winning "Théâtre D'opéra Spatial" by game designer Jason Allen, refined with Midjourney and Photoshop, underscored AI's capacity as a co-creative tool, even securing a prestigious art competition [363]. The commercial viability and artistic acceptance of AI-generated art are further underscored by significant sales at auctions like Christie's "Enhanced Intelligence" in March 2025, where pieces such as Jesse Woolston's "The Dissolution Waiapu" and Holly Herndon & Mat Dryhurst's "Embedding Study 1 & 2" commanded substantial prices [364,365]. Quantitative data from platforms like Pixiv (289,064 AI illustrations/comics)

and ArtStation (15,557 AI search results) further corroborate the widespread adoption of AI tools among visual artists and illustrators [366,367].

In film and animation, generative AI is transitioning from experimental novelty to integrated workflow. The 2025 film "generAIdoscope," entirely conceived and produced using AI for visuals, sound, and score, exemplifies the burgeoning potential for fully AI-driven content creation [368]. More specifically within anime, Frontier Works and KaKa Creation's 2025 experimental series "Twins Hinahima" showcased how Stable Diffusion could significantly reduce costs in the "synthesis and adjustment" phases, also addressing common issues like character clipping in Unreal Engine environments [369].

The gaming sector, a close cousin to anime in visual aesthetics, has also seen considerable AI integration. Nvidia's Deep Learning Super Sampling (DLSS) technology leverages AI for intelligent frame generation, enabling higher resolutions and frame rates without prohibitive hardware upgrades, thereby enhancing player experience [370]. Experimental ventures like Anuttacon's "Whispers From The Star," spearheaded by miHoYo co-founder Cai Haoyu, validate real-time multimodal AI-driven interaction in games. Furthermore, established IPs are incorporating AI; ZUN, the creator of the "Touhou Project," utilized Adobe Firefly for background and texture creation in the 2025 game "Touhou Kinjoukyou Fossilized Wonders [371]." Even independent developers, as seen with the free 2.5D point-and-click adventure game "Echoes of Somewhere," are leveraging AI for generating all in-game art assets, democratizing access to high-quality visuals.

Beyond visual media, AI's influence extends to audio and literature. Pedro Sandoval's 2025 release of two fully AI-generated albums, "ZKY-18" and "Dirty Marilyn," marked a pioneering instance of Spotify-certified AI music. In literature, the 2023 novel "Aum Golly 2," co-created by GPT and Midjourney, represents an emergent form of experimental collaborative authorship.

These diverse examples collectively illustrate that generative AI, including diffusion and language models, is not merely a theoretical construct but a practically adopted and increasingly indispensable set of tools across the creative industries. Their integration, while still evolving, is demonstrably enhancing efficiency, fostering novel artistic expressions, and reshaping traditional production pipelines.

9 Conclusion

This survey has systematically elucidated the transformative impact of diffusion and language models on the landscape of anime generation, charting a trajectory from foundational principles to state-of-the-art applications. Our comprehensive review reveals a field characterized by rapid, albeit uneven, progress. In the domain of image synthesis, models such as Stable Diffusion have achieved considerable maturity, enabling the generation of high-fidelity, stylistically coherent visuals that are already being integrated into artistic workflows. Nevertheless, formidable challenges persist, particularly in maintaining rigorous inter-frame consistency for sequential narratives like manga and mitigating subtle artifacts that detract from professional-grade quality.

The synthesis of higher-dimensional content, particularly video, represents a more nascent yet accelerated frontier. While groundbreaking architectures have significantly enhanced generative fidelity and temporal dynamics, critical bottlenecks remain in achieving robust long-duration spatiotemporal coherence, eliminating visual inconsistencies, and reducing the profound computational exigencies required for training and deployment. Similarly, generative applications in music composition, interactive game creation, and long-form narrative synthesis are still in their emergent phases, grappling with fundamental hurdles related to structural complexity, authentic emotional resonance, semantic integrity, and genuine interactivity.

Looking forward, the continued advancement of this domain is intrinsically predicated upon several key research vectors. The synergistic fusion of autoregressive and diffusion models, particularly through

architectures like Diffusion Transformers (DiTs), presents a promising paradigm for enhancing multimodal coherence and controllability. Future work must also prioritize the development of more computationally efficient paradigms, such as model lightweighting and advanced optimization strategies, to democratize access and facilitate integration into practical production pipelines. Furthermore, the construction of sophisticated world models capable of understanding complex causal and physical relationships will be pivotal for achieving the next echelon of realism and interactivity in generated content. The ultimate objective remains the realization of seamless human-computer collaborative creation, where these technologies function as intuitive and powerful extensions of the human artist's creative intent.

Ultimately, realizing this progressive arc of generative intelligence necessitates a concerted research effort that not only pushes architectural and algorithmic boundaries but also rigorously addresses the attendant ethical, legal, and security implications. Issues of copyright, cultural sensitivity, data bias, and the preservation of human creative autonomy are paramount and demand the establishment of robust governance frameworks. Continuous, critical reflection and interdisciplinary collaboration are imperative to navigate these complexities, ensuring that the profound potential of generative AI is harnessed responsibly to enrich, rather than supplant, the vibrant and evolving art form of anime.

Acknowledgement: The author is grateful for the invaluable support and assistance received from numerous individuals throughout the duration of this project.

Funding Statement: This research was supported by the National Natural Science Foundation of China (Grant No. 62202210).

Author Contributions: Yujie Wu, Xing Deng, and Haijian Shao designed the study. Yujie Wu performed the experiments, collected the data, and conducted the initial data analysis. The entire manuscript was drafted by Yujie Wu, with critical revisions and intellectual contributions from Xing Deng and Haijian Shao. Xing Deng and Haijian Shao provided essential guidance and supervision throughout the project, ensuring the integrity and quality of the research. All authors contributed to the final data analysis and interpretation. Ke Cheng and Ming Zhang provided oversight and contributed to the data analysis. Yingtao Jiang and Fei Wang provided expertise in the conceptualization and guidance of the project, and reviewed and edited the final manuscript. Xing Deng is designated as the corresponding author and guarantor of the study. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Appendix A

Fig. 6: Evolution of Stable Diffusion Models demonstrated via images generated under consistent input conditions.

Prompt: masterpiece, best quality, good quality, very awa, newest, highres, absurdres, 1girl, solo, long hair, looking at viewer, blue eyes, shirt, skirt, black hair, long sleeves, closed mouth, jewelry, sitting, white shirt, outdoors, sky, barefoot, cloud, water, from side, bracelet, ocean, scenery, reflection, rock, ruins, bridge, power lines, river, utility pole, evening, reflective water, rubble, limited palette, sketch, glowing, psychedelic, epic composition, epic proportion, dynamic angle, volumetric lighting, masterpiece, best quality, amazing quality, very aesthetic, absurdres, newest, perfect hands, OverCute, FlatLineArt

Checkpoints Models: Stable Diffusion 1.5, Stable Diffusion XL, Stable Diffusion 2.0, Stable Diffusion 3.0, Stable Diffusion 3.5, FLUX 1.0

Hardware: TensorArt Cloud Generation Service

Fig. 9: Depictions generated by various contemporary large-scale generative models using a consistent textual prompt and parameters.

Prompt: Smooth Quality, 1girl, solo, mature, parted lips, instrument, microphone, meteor shower, starry sky, music, guitar, holding instrument, electric guitar, facing down, messy hair, crazy hair, from below, cowboy shot, epic composition, epic proportion, dynamic angle, masterpiece, best quality, amazing quality, very aesthetic, absurdres, newest, perfect hands, FlatLineArt

Dataset Size: 50 Images

Checkpoints Models: Firefly, CogView3, ImagenFX, Lumina2, HunyuanDit, Ideogram, Illustrious (SDXL), Pony (SDXL), Stable Diffusion 3.5, Niji_style (SD3.5), Flux1, ToxicEchoFlux (FLUX)

Image Generation Hardware: TensorArt Cloud Generation Service, API services provided by relevant commercial companies

Image Testing Hardware: NVIDIA GeForce RTX 4090

Figs. 11–18

Positive Prompt: Neptune-Hyperdimension Neptunia, 4K FlatLineArt, digital drawing mode, an energetic young girl with light purple hair styled in pigtails and bright purple eyes, wearing her signature purple and white outfit, walking towards the camera from a distance on a brightly lit city street at daytime, full body, dynamic and playful stance, cinematic lighting, perfect anatomy, detailed outfit and D-pad hair ornaments, full HD, 4K, HDR, depth of field

Negative Prompt: bright colors, overexposed, static, blurred details, subtitles, style, artwork, painting, picture, still, overall gray, worst quality, low quality, JPEG compression residue, ugly, incomplete, extra fingers, poorly drawn hands, poorly drawn faces, deformed, disfigured, malformed limbs, fused fingers, still picture, cluttered background, three legs, many people in the background, walking backwards

Checkpoints Models: CogVideoX1.5 5B, Cosmos-1.0-7B, HunyuanVideo, LTX-Video 2B, Mochi 1 10B, Pyramid-Flow-miniFlux, SkyReels-V1, WAN_2_1

Video Generation Hardware: TensorArt Cloud Generation Service, API services provided by relevant commercial companies

Video Testing Hardware: NVIDIA GeForce RTX 4090

References

1. Chen Y, Liu C, Huang W, Cheng S, Arcucci R, Xiong Z. Generative text-guided 3D vision-language pretraining for unified medical image segmentation. arXiv:2306.04811. 2023.
2. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. arXiv:2006.11239. 2020.
3. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE; 2022. p. 10684–95. [cited 2025 Apr 8]. Available from: https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.
4. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. arXiv:1406.2661. 2014.
5. Dhariwal P, Nichol A. Classifier guidance: diffusion models beat GANs on image synthesis. arXiv:2105.05233. 2021.

6. Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, et al. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. arXiv:2112.10741. 2022 [cited 2025 Apr 11]. Available from: http://arxiv.org/abs/2112.10741.
7. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. 2019 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/1810.04805.
8. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training [Internet]. OpenAI; 2018 Jun 11 [cited 2025 Apr 8]. 14 p. Available from: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
9. Cao Y, Meng X, Mok PY, Lee TY, Liu X, Li P. AnimeDiffusion: anime diffusion colorization. IEEE Trans Vis Comput Graph. 2024;30(10):6956–69. doi:10.1109/tvcg.2024.3357568.
10. Yang M. Research on anime-style image generation based on stable diffusion. ITM Web Conf. 2025;73(4):02038. doi:10.1051/itmconf/20257302038.
11. Wagan SM, Sidra S. Revolutionizing the digital creative industries: the role of artificial intelligence in integration, development, and innovation. SEISENSE J Manag. 2024;7(1):135–52. doi:10.33215/rvcwy166.
12. Schuhmann C, Beaumont R, Vencu R, Gordon C, Wightman R, Cherti M, et al. LAION-5B: an open large-scale dataset for training next generation image-text models. arXiv:2210.08402. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2210.08402.
13. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc.; 2017 [cited 2025 Apr 8]. Available from: https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.
14. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. Cambridge, MA, USA: PMLR; 2021 [cited 2025 Apr 8]. p. 8748–63. Available from: https://proceedings.mlr.press/v139/radford21a.html.
15. Amankwah-Amoah J, Abdalla S, Mogaji E, Elbanna A, Dwivedi YK. The impending disruption of creative industries by generative AI: opportunities, challenges, and research agenda. Int J Inf Manag. 2024;79(2):102759. doi:10.1016/j.ijinfomgt.2024.102759.
16. Tang MY, Chen Y. AI and animated character design: efficiency, creativity, interactivity. Fronti Soc Sci Technol. 2024;6(1). doi:10.25236/fsst.2024.060120.
17. Chen Y, Wang Y, Yu T, Pan Y. The effect of AI on animation production efficiency: an empirical investigation through the network data envelopment analysis. Electronics. 2024;13(24):5001. doi:10.3390/electronics13245001.
18. Salimans T, Ho J. Progressive distillation for fast sampling of diffusion models. arXiv:2202.00512. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2202.00512.
19. Nichol A, Dhariwal P. Improved denoising diffusion probabilistic models. arXiv:2102.09672. 2021 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2102.09672.
20. Song Y, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, Poole B. Score-based generative modeling through stochastic differential equations. arXiv:2011.13456. 2021 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2011.13456.
21. Karras T, Aittala M, Aila T, Laine S. Elucidating the design space of diffusion-based generative models. Adv Neural Inf Process Syst. 2022;35:26565–77.
22. Song J, Meng C, Ermon S. Denoising diffusion implicit models. arXiv:2010.02502. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2010.02502.
23. Hoogeboom E, Nielsen D, Jaini P, Forré P, Welling M. Argmax flows and multinomial diffusion: learning categorical distributions. arXiv:2102.05379. 2021 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2102.05379.
24. Austin J, Johnson DD, Ho J, Tarlow D, van den Berg R. Structured denoising diffusion models in discrete state-spaces. arXiv:2107.03006. 2023 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2107.03006.
25. Gu S, Chen D, Bao J, Wen F, Zhang B, Chen D, et al. Vector quantized diffusion model for text-to-image synthesis. arXiv:2111.14822. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2111.14822.

26. Liu L, Ren Y, Lin Z, Zhao Z. Pseudo numerical methods for diffusion models on manifolds. arXiv:2202.09778. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2202.09778.
27. Bao F, Li C, Zhu J, Zhang B. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. arXiv:2201.06503. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2201.06503.
28. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical text-conditional image generation with CLIP latents. arXiv:2204.06125. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2204.06125.
29. Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, et al. Photorealistic text-to-image diffusion models with deep language understanding. Adv Neural Inform Process Syst. 2022 Dec 6;35:36479–94.
30. Pokle A, Geng Z, Kolter Z. Deep equilibrium approaches to diffusion models. arXiv:2210.12867. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2210.12867.
31. Guo L, Wang C, Yang W, Huang S, Wang Y, Pfister H, et al. ShadowDiffusion: when degradation prior meets diffusion model for shadow removal. arXiv:2212.04711. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2212.04711.
32. Poole B, Jain A, Barron JT, Mildenhall B. DreamFusion: text-to-3D using 2D diffusion. arXiv:2209.14988. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2209.14988.
33. Lin CH, Gao J, Tang L, Takikawa T, Zeng X, Huang X, et al. Magic3D: high-resolution text-to-3D content creation. arXiv:2211.10440. 2023 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2211.10440.
34. He Y, Yang T, Zhang Y, Shan Y, Chen Q. Latent video diffusion models for high-fidelity long video generation. arXiv:2211.13221. 2023 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2211.13221.
35. Wu J, Fu R, Fang H, Zhang Y, Yang Y, Xiong H, et al. MedSegDiff: medical image segmentation with diffusion probabilistic model. arXiv:2211.00611. 2023 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2211.00611.
36. Singer U, Polyak A, Hayes T, Yin X, An J, Zhang S, et al. Make-a-video: text-to-video generation without text-video data. arXiv:2209.14792. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2209.14792.
37. Ho J, Salimans T, Gritsenko A, Chan W, Norouzi M, Fleet DJ. Video diffusion models. arXiv:2204.03458. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2204.03458.
38. Nichol A, Jun H, Dhariwal P, Mishkin P, Chen M. Point-E: a system for generating 3D point clouds from complex prompts. arXiv:2212.08751. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2212.08751.
39. Wang Y, Ren S, Lin Z, Han Y, Guo H, Yang Z, et al. Parallelized autoregressive visual generation. arXiv:2412.15119. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.15119.
40. Li XL, Thickstun J, Gulrajani I, Liang P, Hashimoto TB. Diffusion-LM improves controllable text generation. arXiv:2205.14217. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2205.14217.
41. Changpinyo S, Sharma P, Ding N, Soricut R. Conceptual 12M: pushing web-scale image-text pre-training to recognize long-tail visual concepts. arXiv:2102.08981. 2021 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2102.08981.
42. Li J, Selvaraju RR, Gotmare AD, Joty S, Xiong C, Hoi S. Align before Fuse: vision and language representation learning with momentum distillation. arXiv:2107.07651. 2021 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2107.07651.
43. Rae JW, Borgeaud S, Cai T, Millican K, Hoffmann J, Song F, et al. Scaling language models: methods, analysis & insights from training gopher. arXiv:2112.11446. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2112.11446.
44. Podell D, English Z, Lacey K, Blattmann A, Dockhorn T, Müller J, et al. SDXL: improving latent diffusion models for high-resolution image synthesis. arXiv:2307.01952. 2023 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2307.01952.
45. Bain M, Nagrani A, Varol G, Zisserman A. Frozen in time: a joint video and image encoder for end-to-end retrieval. arXiv:2104.00650. 2022 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2104.00650.
46. Xue H, Hang T, Zeng Y, Sun Y, Liu B, Yang H, et al. Advancing high-resolution video-language representation with large-scale video transcriptions. arXiv:2111.10337. 2022 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2111.10337.

47. Wang W, Yang Y. VidProM: a million-scale real prompt-gallery dataset for text-to-video diffusion models. arXiv:2403.06098. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2403.06098.
48. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20–25; Miami, FL, USA. p. 248–55.
49. Lin TY, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, et al. Microsoft COCO: common objects in context. arXiv:1405.0312. 2015 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/1405.0312.
50. Liu Z, Luo P, Wang X, Tang X. Deep learning face attributes in the wild. arXiv:1411.7766. 2015 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/1411.7766.
51. Jayasumana S, Ramalingam S, Veit A, Glasner D, Chakrabarti A, Kumar S, et al. Rethinking FID: towards a better evaluation metric for image generation. Jun 5]3. 2024 [cited 2025 Jun 5]. Available from: http://arxiv.org/abs/2401.09603.
52. Peebles W, Xie S. Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE/CVF; 2023 [cited 2025 Apr 8]. p. 4195–205. Available from: https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.
53. Stability AI. Stable diffusion 3: research paper. [cited 2025 Apr 8]. Available from: https://stability.ai/news/stable-diffusion-3-research-paper.
54. Ho J, Saharia C, Chan W, Fleet DJ, Norouzi M, Salimans T. Cascaded diffusion models for high fidelity image generation. arXiv:2106.15282. 2021 [cited 2025 Jun 5]. Available from: http://arxiv.org/abs/2106.15282.
55. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners [Internet]. OpenAI; 2019 Feb 14 [cited 2025 Apr 8]. 13 p. Available from: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
56. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692. 2019 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/1907.11692.
57. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1–67.
58. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc.; 2020 [cited 2025 Apr 8]. p. 1877–901. Available from: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.
59. Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, et al. The Pile: an 800GB dataset of diverse text for language modeling. arXiv:2101.00027. 2020 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2101.00027.
60. Thoppilan R, Freitas DD, Hall J, Shazeer N, Kulshreshtha A, Cheng HT, et al. LaMDA: language models for dialog applications. arXiv:2201.08239. 2022 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2201.08239.
61. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: scaling language modeling with pathways. arXiv:2204.02311. 2022 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2204.02311.
62. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. arXiv:2303.08774. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2303.08774.
63. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. arXiv:2302.13971. 2023 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2302.13971.
64. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc.; 2017 [cited 2025 Apr 8]. Available from: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
65. Du Z, Qian Y, Liu X, Ding M, Qiu J, Yang Z, et al. GLM: general language model pretraining with autoregressive blank infilling. arXiv:2103.10360. 2022 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2103.10360.
66. Lynch CJ, Jensen E, Munro MH, Zamponi V, Martinez J, O'Brien K, et al. GPT-4 generated narratives of life events using a structured narrative prompt: a validation study. arXiv:2402.05435. 2024 [cited 2025 Apr 13]. Available from: http://arxiv.org/abs/2402.05435.

67. Su J, Lu Y, Pan S, Murtadha A, Wen B, Liu Y. RoFormer: enhanced transformer with rotary position embedding. arXiv:2104.09864. 2023 [cited 2025 Jun 5]. Available from: http://arxiv.org/abs/2104.09864.
68. Press O, Smith NA, Lewis M. Train short, test long: attention with linear biases enables input length extrapolation. arXiv:2108.12409. 2022 [cited 2025 Jun 5]. Available from: http://arxiv.org/abs/2108.12409.
69. Kitaev N, Kaiser Ł., Levskaya A. Reformer: the efficient transformer. arXiv:2001.04451. 2020 [cited 2025 Jun 5]. Available from: http://arxiv.org/abs/2001.04451.
70. Zaheer M, Guruganesh G, Dubey A, Ainslie J, Alberti C, Ontanon S, et al. Big bird: transformers for longer sequences. arXiv:2007.14062. 2021 [cited 2025 Jun 5]. Available from: http://arxiv.org/abs/2007.14062.
71. Dao T, Fu DY, Ermon S, Rudra A, Ré C. FlashAttention: fast and memory-efficient exact attention with IO-awareness. arXiv:2205.14135. 2022 [cited 2025 Jun 5]. Available from: http://arxiv.org/abs/2205.14135.
72. Han Z, Gao C, Liu J, Zhang J, Zhang SQ. Parameter-efficient fine-tuning for large models: a comprehensive survey. arXiv:2403.14608. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2403.14608.
73. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. arXiv:2106.09685. 2021 [cited 2025 Apr 11]. Available from: http://arxiv.org/abs/2106.09685.
74. Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, Laroussilhe QD, Gesmundo A, et al. Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning; 2019 Jun 9–15; Long Beach, CA, USA; 2019 [cited 2025 May 21]. p. 2790–9. Available from: https://proceedings.mlr.press/v97/houlsby19a.html.
75. Li XL, Liang P. Prefix-tuning: optimizing continuous prompts for generation. arXiv:2101.00190. 2021 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2101.00190.
76. Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. arXiv:2104.08691. 2021 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2104.08691.
77. Wu Z, Arora A, Wang Z, Geiger A, Jurafsky D, Manning CD, et al. ReFT: representation finetuning for language models. arXiv:2404.03592. 2024 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2404.03592.
78. Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 3813–24.
79. Ruiz N, Li Y, Jampani V, Pritch Y, Rubinstein M, Aberman K. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation. arXiv:2208.12242. 2023 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2208.12242.
80. Gal R, Alaluf Y, Atzmon Y, Patashnik O, Bermano AH, Chechik G, et al. An image is worth one word: personalizing text-to-image generation using textual inversion. arXiv:2208.01618. 2022 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2208.01618.
81. Ha D, Dai A, Le QV. HyperNetworks. arXiv:1609.09106. 2016 [cited 2025 May 21]. Available from: http://arxiv.org/abs/1609.09106.
82. Davila A, Colan J, Hasegawa Y. Comparison of fine-tuning strategies for transfer learning in medical image classification. Image Vis Comput. 2024;146(2):105012. doi:10.1016/j.imavis.2024.105012.
83. Parthasarathy VB, Zafar A, Khan A, Shahid A. The ultimate guide to fine-tuning LLMs from basics to breakthroughs: an exhaustive review of technologies, research, best practices, applied research challenges and opportunities. arXiv:2408.13296. 2024 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2408.13296.
84. Lippmann P, Skublicki K, Tanner J, Ishiwatari S, Yang J. Context-Informed machine translation of manga using multimodal large language models. arXiv:2411.02589. 2024 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2411.02589.
85. Sachdeva R, Shin G, Zisserman A. Tails tell tales: chapter-wide manga transcriptions with character names. arXiv:2408.00298. 2024 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2408.00298.
86. Vivoli E, Bertini M, Karatzas D. CoMix: a comprehensive benchmark for multi-task comic understanding. arXiv:2407.03550. 2024 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2407.03550.
87. Ju X, Zeng A, Wang J, Xu Q, Zhang L. Human-Art: a versatile human-centric dataset bridging natural and artificial scenes. arXiv:2303.02760. 2023 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2303.02760.

88. Li Y, Aizawa K, Matsui Y. Manga109Dialog: a large-scale dialogue dataset for comics speaker detection. arXiv:2306.17469. 2024 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2306.17469.
89. Seo CW, Ashtari A, Noh J. Semi-supervised reference-based sketch extraction using a contrastive learning framework. arXiv:2407.14026. 2024 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2407.14026.
90. Li Z, Xu Y, Zhao N, Zhou Y, Liu Y, Lin D, et al. Parsing-conditioned anime translation: a new dataset and method. ACM Trans Graph. 2023;42(3):30:1–30:14. doi:10.1145/3585002.
91. Siyao L, Li Y, Li B, Dong C, Liu Z, Loy CC. AnimeRun: 2D animation visual correspondence from open source 3D movies. arXiv:2211.05709. 2022 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2211.05709.
92. Baek J, Matsui Y, Aizawa K. COO: comic onomatopoeia dataset for recognizing arbitrary or truncated texts. arXiv:2207.04675. 2022 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2207.04675.
93. Kim K, Park S, Lee J, Chung S, Lee J, Choo J. AnimeCeleb: large-scale animation celebheads dataset for head reenactment. arXiv:2111.07640. 2022 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2111.07640.
94. Rios EA, Cheng WH, Lai BC. DAF: re: a challenging, crowd-sourced, large-scale, long-tailed dataset for anime character recognition. arXiv:2101.08674. 2021 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2101.08674.
95. Zheng Y, Zhao Y, Ren M, Yan H, Lu X, Liu J, et al. Cartoon face recognition: a benchmark dataset. arXiv:1907.13394. 2020 [cited 2025 May 21]. Available from: http://arxiv.org/abs/1907.13394.
96. Gobbo JD, Herrera RM. Unconstrained text detection in manga: a new dataset and baseline. arXiv:2009.04042. 2020 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2009.04042.
97. Zhang L, Ji Y, Liu C. An illustration region dataset. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Computer vision—ECCV 2020, Lecture notes in computer science. Vol. 12358. Cham: Springer International Publishing; 2020 [cited 2025 May 21]. p. 137–54. Available from: https://link.springer.com/10.1007/978-3-030-58601-0_9.
98. Aizawa K, Fujimoto A, Otsubo A, Ogawa T, Matsui Y, Tsubota K, et al. Building a manga dataset "Manga109" with annotations for multimedia applications. IEEE Multimed. 2020;27(2):8–18. doi:10.1109/mmul.2020.2987895.
99. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of StyleGAN. arXiv:1912.04958. 2020 [cited 2025 May 21]. Available from: http://arxiv.org/abs/1912.04958.
100. Shi X, Hu W. A survey of image style transfer research. In: 2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI); 2022 Sep 23–25; Nanjing, China; 2022 [cited 2025 May 21]. p. 133–7. Available from: https://ieeexplore.ieee.org/document/9950226.
101. Li S, Tang H. Multimodal alignment and fusion: a survey. arXiv:2411.17040. 2024 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2411.17040.
102. Olamendy JC. Tackling the challenge of imbalanced datasets: a comprehensive guide. Medium. 2024. [cited 2025 May 21]. Available from: https://medium.com/@juanc.olamendy/tackling-the-challenge-of-imbalanced-datasets-a-comprehensive-guide-2feb11ca2fa0.
103. black-forest-labs/flux. black-forest-labs; 2025 [cited 2025 Jul 21]. Available from: https://github.com/black-forest-labs/flux.
104. Midjourney; 2025 [cited 2025 Jul 20]. Updates. Available from: https://www.midjourney.com/website.
105. OpenAI Help Center. [cited 2025 Jul 20]. ChatGPT—Release Notes. Available from: https://help.openai.com/en/articles/6825453-chatgpt-release-notes#h_8670d7da97.
106. Adobe Firefly-Free Generative AI for creatives. [cited 2025 Apr 12]. Available from: https://www.adobe.com/products/firefly.html.
107. Ideogram 3.0. [cited 2025 Jul 20]. Available from: https://about.ideogram.ai/3.0.
108. THUDM/CogView4. Z.ai & THUKEG; 2025 [cited 2025 Jul 20]. Available from: https://github.com/THUDM/CogView4.
109. Kwai-Kolors/Kolors. Kolors; 2025 [cited 2025 Jul 21]. Available from: https://github.com/Kwai-Kolors/Kolors.
110. Gao Y, Gong L, Guo Q, Hou X, Lai Z, Li F, et al. Seedream 3.0 technical report. arXiv:2504.11346. 2025 [cited 2025 Aug 17]. Available from: http://arxiv.org/abs/2504.11346.
111. Li Z, Zhang J, Lin Q, Xiong J, Long Y, Deng X, et al. Hunyuan-DiT: a powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv:2405.08748. 2024 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2405.08748.

112. Flux AI-Free Online Advanced Flux AI Image Generator. [cited 2025 Apr 11]. Available from: https://flux1.ai/.
113. Mou C, Wang X, Xie L, Wu Y, Zhang J, Qi Z, et al. T2I-Adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv:2302.08453. 2023 [cited 2025 Apr 11]. Available from: http://arxiv.org/abs/2302.08453.
114. Kawano Y, Aoki Y. MaskDiffusion: exploiting pre-trained diffusion models for semantic segmentation. arXiv:2403.11194. 2024 [cited 2025 Apr 11]. Available from: http://arxiv.org/abs/2403.11194.
115. Luo S, Tan Y, Huang L, Li J, Zhao H. Latent consistency models: synthesizing high-resolution images with few-step inference. arXiv:2310.04378. 2023 [cited 2025 Apr 11]. Available from: http://arxiv.org/abs/2310.04378.
116. AUTOMATIC1111. AUTOMATIC1111/stable-diffusion-webui-feature-showcase. 2025 [cited 2025 Apr 11]. Available from: https://github.com/AUTOMATIC1111/stable-diffusion-webui-feature-showcase.
117. comfyanonymous. comfyanonymous/ComfyUI. 2025 [cited 2025 Apr 11]. Available from: https://github.com/comfyanonymous/ComfyUI.
118. lllyasviel. lllyasviel/Paints-UNDO. 2025 [cited 2025 Apr 11]. Available from: https://github.com/lllyasviel/Paints-UNDO.
119. Imagen-Team-Google, Baldridge J, Bauer J, Bhutani M, Brichtova N, Bunner A, et al. Imagen 3. arXiv:2408.07009. 2024 [cited 2025 Apr 11]. Available from: http://arxiv.org/abs/2408.07009.
120. Team G, Mesnard T, Hardin C, Dadashi R, Bhupatiraju S, Pathak S, et al. Gemma: open models based on gemini research and technology. arXiv:2403.08295. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2403.08295.
121. Qin Q, Zhuo L, Xin Y, Du R, Li Z, Fu B, et al. Lumina-Image 2.0: a unified and efficient image generative framework. arXiv:2503.21758. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2503.21758.
122. Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller M. Deterministic policy gradient algorithms. In: Proceedings of the 31st International Conference on Machine Learning. Bejing, China: PMLR; 2014. Available from: https://proceedings.mlr.press/v32/silver14.html.
123. Ghosh D, Hajishirzi H, Schmidt L. GenEval: an object-focused framework for evaluating text-to-image alignment. arXiv:2310.11513. 2023 [cited 2025 Apr 11]. Available from: http://arxiv.org/abs/2310.11513.
124. Zheng W, Teng J, Yang Z, Wang W, Chen J, Gu X, et al. CogView3: finer and faster text-to-image generation via relay diffusion. arXiv:2403.05121. 2024 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2403.05121.
125. Betker J, Goh G, Jing L, Brooks T, Wang J, Li L, et al. Improving image generation with better captions [Internet]. OpenAI; 2023. Available from: https://cdn.openai.com/papers/dall-e-3.pdf.
126. Park SH, Koh JY, Lee J, Song J, Kim D, Moon H, et al. Illustrious: an open advanced illustration model. arXiv:2409.19946. 2024 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2409.19946.
127. Pony Diffusion V6 XL - V6 (start with this one) | Stable Diffusion Checkpoint | Civitai. 2025 [cited 2025 Apr 12]. Available from: https://civitai.com/models/257749/pony-diffusion-v6-xl.
128. Sehwag V, Kong X, Li J, Spranger M, Lyu L. Stretching each dollar: diffusion training from scratch on a micro-budget. arXiv:2407.15811. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2407.15811.
129. Xie E, Chen J, Chen J, Cai H, Tang H, Lin Y, et al. SANA: efficient high-resolution image synthesis with linear diffusion transformers. arXiv:2410.10629. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2410.10629.
130. Mi Z, Wang KC, Qian G, Ye H, Liu R, Tulyakov S, et al. I think, therefore i diffuse: enabling multimodal in-context reasoning in diffusion models. arXiv:2502.10458. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2502.10458.
131. Chen L, Bai S, Chai W, Xie W, Zhao H, Vinci L, et al. Multimodal representation alignment for image generation: text-image interleaved control is easier than you think. arXiv:2502.20172. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2502.20172.
132. Liu Z, Cheng KL, Chen X, Xiao J, Ouyang H, Zhu K, et al. MangaNinja: line art colorization with precise reference following. arXiv:2501.08332. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.08332.
133. Huang S, Song Y, Zhang Y, Guo H, Wang X, Shou MZ, et al. PhotoDoodle: learning artistic image editing from few-shot pairwise data. arXiv:2502.14397. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2502.14397.

134. Li J, Cao J, Guo Y, Li W, Zhang Y. One diffusion step to real-world super-resolution via flow trajectory distillation. arXiv:2502.01993. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2502.01993.
135. Sun Z, Xiong S, Chen Y, Rong Y. Content-style decoupling for unsupervised makeup transfer without generating pseudo ground truth. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 7601–10.
136. Ye H, Zhang J, Liu S, Han X, Yang W. IP-Adapter: text compatible image prompt adapter for text-to-image diffusion models. arXiv:2308.06721. 2023 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2308.06721.
137. Wang Q, Bai X, Wang H, Qin Z, Chen A, Li H, et al. InstantID: zero-shot identity-preserving generation in seconds. arXiv:2401.07519. 2024 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2401.07519.
138. Li Z, Cao M, Wang X, Qi Z, Cheng MM, Shan Y. PhotoMaker: customizing realistic human photos via stacked ID embedding. arXiv:2312.04461. 2023 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2312.04461.
139. Zhou Y, Zhou D, Cheng MM, Feng J, Hou Q. StoryDiffusion: consistent self-attention for long-range image and video generation. Adv Neural Inf Process Syst. 2024;37:110315–40.
140. Wu J, Tang C, Wang J, Zeng Y, Li X, Tong Y. DiffSensei: bridging multi-modal LLMs and diffusion models for customized manga generation. arXiv:2412.07589. 2025 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2412.07589.
141. He J, Tuo Y, Chen B, Zhong C, Geng Y, Bo L. AnyStory: towards unified single and multiple subject personalization in text-to-image generation. arXiv:2501.09503. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.09503.
142. Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. arXiv:1809.11096. 2019 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/1809.11096.
143. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 2004;13(4):600–12. doi:10.1109/tip.2003.819861.
144. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. arXiv:1606.03498. 2016 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/1606.03498.
145. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 20–23; Salt Lake City, UT, USA. p. 586–95.
146. Bińkowski M, Sutherland DJ, Arbel M, Gretton A. Demystifying MMD GANs. arXiv:1801.01401. 2021 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/1801.01401.
147. Liu Y, Zhang K, Li Y, Yan Z, Gao C, Chen R, et al. Sora: a review on background, technology, limitations, and opportunities of large vision models. arXiv:2402.17177. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2402.17177.
148. team TVPU. Veo product updates. 2025 [cited 2025 Jul 21]. Available from: https://updates.veo.co/.
149. Polyak A, Zohar A, Brown A, Tjandra A, Sinha A, Lee A, et al. Movie gen: a cast of media foundation models. arXiv:2410.13720. 2025 [cited 2025 Jul 21]. Available from: http://arxiv.org/abs/2410.13720.
150. HaCohen Y, Chiprut N, Brazowski B, Shalem D, Moshe D, Richardson E, et al. LTX-Video: realtime video latent diffusion. arXiv:2501.00103. 2024 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2501.00103.
151. genmoai/mochi. Genmo; 2025 [cited 2025 Jul 21]. Available from: https://github.com/genmoai/mochi.
152. Hong W, Ding M, Zheng W, Liu X, Tang J. CogVideo: large-scale pretraining for text-to-video generation via transformers. arXiv:2205.15868. 2022 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2205.15868.
153. Kong W, Tian Q, Zhang Z, Min R, Dai Z, Zhou J, et al. HunyuanVideo: a systematic framework for large video generative models. arXiv:2412.03603. 2025 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2412.03603.
154. NNVIDIA, Agarwal N, Ali A, Bala M, Balaji Y, Barker E, et al. Cosmos world foundation model platform for physical AI. arXiv:2501.03575. 2025 [cited 2025 Aug 17]. Available from: http://arxiv.org/abs/2501.03575.
155. Qiu D, Fei Z, Wang R, Bai J, Yu C, Fan M, et al. SkyReels-A1: expressive portrait animation in video diffusion transformers. arXiv:2502.10841. 2025 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2502.10841.
156. WanTeam, Wang A, Ai B, Wen B, Mao C, Xie CW, et al. Wan: open and advanced large-scale video generative models. arXiv:2503.20314. 2025 [cited 2025 Apr 18]. Available from: http://arxiv.org/abs/2503.20314.

157. Runway | Tools for human imagination. [cited 2025 Jul 21]. Available from: https://runwayml.com/changelog.
158. pikalabsorg. Pika 2.2. Pika Labs. 2025 [cited 2025 Jul 21]. Available from: https://pikalabs.org/pika-2-2/.
159. Luma AI | AI Video Generation with Ray2 & Dream Machine | Luma AI. [cited 2025 Jul 21]. Available from: https://lumalabs.ai/.
160. Blattmann A, Dockhorn T, Kulal S, Mendelevitch D, Kilian M, Lorenz D, et al. Stable video diffusion: scaling latent video diffusion models to large datasets. arXiv:2311.15127. 2023 [cited 2025 May 21]. Available from: http://arxiv.org/abs/2311.15127.
161. Kling AI: Next-Gen AI Video & AI Image Generator. [cited 2025 Jul 21]. Available from: https://app.klingai.com/global/.
162. hailuoai.video. Hailuo AI: Transform Idea to Visual with AI–AI Generated Video | Hailuo AI. [cited 2025 Aug 17]. Available from: https://hailuoai.video.
163. Jin Y, Sun Z, Li N, Xu K, Xu K, Jiang H, et al. Pyramidal flow matching for efficient video generative modeling. arXiv:2410.05954. 2025 [cited 2025 Jul 21]. Available from: http://arxiv.org/abs/2410.05954.
164. Lei J, Hu X, Wang Y, Liu D. PyramidFlow: high-resolution defect contrastive localization using pyramid normalizing flow. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada: IEEE. p. 14143–52.
165. Liu R, Wu H, Ziqiang Z, Wei C, He Y, Pi R, et al. VideoDPO: omni-preference alignment for video diffusion generation. arXiv:2412.14167. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.14167.
166. Zhang DJ, Wu JZ, Liu JW, Zhao R, Ran L, Gu Y, et al. Show-1: marrying pixel and latent diffusion models for text-to-video generation. Int J Comput Vis. 2025;133(4):1879–93. doi:10.1007/s11263-024-02271-9.
167. Chen H, Zhang Y, Cun X, Xia M, Wang X, Weng C, et al. VideoCrafter2: overcoming data limitations for high-quality video diffusion models. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 7310–20.
168. Huang Y, Zheng W, Gao Y, Tao X, Wan P, Zhang D, et al. Owl-1: omni world model for consistent long video generation. arXiv:2412.09600. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.09600.
169. Wu Z, Siarohin A, Menapace W, Skorokhodov I, Fang Y, Chordia V, et al. Mind the time: temporally-controlled multi-event video generation. arXiv:2412.05263. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.05263.
170. Yin S, Wu C, Yang H, Wang J, Wang X, Ni M, et al. NUWA-XL: diffusion over diffusion for eXtremely long video generation. arXiv:2303.12346. 2023 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2303.12346.
171. Yan X, Cai Y, Wang Q, Zhou Y, Huang W, Yang H. Long video diffusion generation with segmented cross-attention and content-rich video data curation. arXiv:2412.01316. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.01316.
172. Atzmon Y, Gal R, Tewel Y, Kasten Y, Chechik G. Motion by queries: identity-motion trade-offs in text-to-video generation. arXiv:2412.07750. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.07750.
173. Xu Z, Zhang J, Liew JH, Yan H, Liu JW, Zhang C, et al. MagicAnimate: temporally consistent human image animation using diffusion model. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 1481–90.
174. Lei G, Wang C, Li H, Zhang R, Wang Y, Xu W. AnimateAnything: consistent and controllable animation for video generation. arXiv:2411.10836. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2411.10836.
175. Gal R, Vinker Y, Alaluf Y, Bermano A, Cohen-Or D, Shamir A, et al. Breathing life into sketches using text-to-video priors. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 4325–36.
176. Liu G, Xia M, Zhang Y, Chen H, Xing J, Wang Y, et al. StyleCrafter: enhancing stylized text-to-video generation with style adapter. arXiv:2312.00330. 2024 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2312.00330.
177. Yao J, Yang B, Wang X. Reconstruction vs. generation: taming optimization dilemma in latent diffusion models. arXiv:2501.01423. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.01423.

178. Wang Q, Luo Y, Shi X, Jia X, Lu H, Xue T, et al. CineMaster: a 3D-aware and controllable framework for cinematic text-to-video generation. arXiv:2502.08639. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2502.08639.
179. Hyung J, Kim K, Hong S, Kim MJ, Choo J. Spatiotemporal skip guidance for enhanced video diffusion sampling. arXiv:2411.18664. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2411.18664.
180. Liu L, Ma T, Li B, Chen Z, Liu J, He Q, et al. Phantom: subject-consistent video generation via cross-modal alignment. arXiv:2502.11079. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2502.11079.
181. Zhuang S, Li K, Chen X, Wang Y, Liu Z, Qiao Y, et al. Vlogger: make your dream a vlog. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 8806–17.
182. Long F, Qiu Z, Yao T, Mei T. VideoStudio: generating consistent-content and multi-scene videos. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer vision–ECCV 2024. Cham: Springer Nature Switzerland; 2025. p. 468–85. doi:10.1007/978-3-031-73027-6_27.
183. Liu X, Zeng A, Xue W, Yang H, Luo W, Liu Q, et al. VFX creator: animated visual effect generation with controllable diffusion transformer. arXiv:2502.05979. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2502.05979.
184. Xing J, Xia M, Zhang Y, Chen H, Yu W, Liu H, et al. DynamiCrafter: animating open-domain images with video diffusion priors. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer vision–ECCV 2024. Cham: Springer Nature Switzerland; 2025. p. 399–417. doi:10.1007/978-3-031-72952-2_23.
185. Zhang S, Wang J, Zhang Y, Zhao K, Yuan H, Qin Z, et al. I2VGen-XL: high-quality image-to-video synthesis via cascaded diffusion models. arXiv:2311.04145. 2023 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2311.04145.
186. Cui J, Li H, Zhan Y, Shang H, Cheng K, Ma Y, et al. Hallo3: highly dynamic and realistic portrait image animation with video diffusion transformer. arXiv:2412.00733. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.00733.
187. Liu T, Ma Z, Chen Q, Chen F, Fan S, Chen X, et al. VQTalker: towards multilingual talking avatars through facial motion tokenization. arXiv:2412.09892. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.09892.
188. Zheng L, Zhang Y, Guo H, Pan J, Tan Z, Lu J, et al. MEMO: memory-guided diffusion for expressive talking video generation. arXiv:2412.04448. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.04448.
189. Li C, Zhang C, Xu W, Lin J, Xie J, Feng W, et al. LatentSync: taming audio-conditioned latent diffusion models for lip sync with SyncNet supervision. arXiv:2412.09262. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.09262.
190. Zhu Y, Zhang L, Rong Z, Hu T, Liang S, Ge Z. INFP: audio-driven interactive head generation in dyadic conversations. arXiv:2412.04037. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.04037.
191. Cui J, Li H, Yao Y, Zhu H, Shang H, Cheng K, et al. Hallo2: long-duration and high-resolution audio-driven portrait image animation. arXiv:2410.07718. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2410.07718.
192. Tian L, Hu S, Wang Q, Zhang B, Bo L. EMO2: end-effector guided audio-driven avatar video generation. arXiv:2501.10687. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.10687.
193. Zhai Y, Lin K, Li L, Lin CC, Wang J, Yang Z, et al. IDOL: unified dual-modal latent diffusion for human-centric joint video-depth generation. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer vision–ECCV 2024. Cham: Springer Nature Switzerland; 2025. p. 134–52. doi:10.1007/978-3-031-72633-0_8.
194. Zhu S, Chen JL, Dai Z, Dong Z, Xu Y, Cao X, et al. Champ: controllable and consistent human image animation with 3D parametric guidance. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer vision–ECCV 2024. Cham: Springer Nature Switzerland; 2025. p. 145–62. doi:10.1007/978-3-031-73001-6_9.
195. Liu K, Shao L, Lu S. Novel view extrapolation with video diffusion priors. arXiv:2411.14208. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2411.14208.
196. Liu S, Ren Z, Gupta S, Wang S. PhysGen: rigid-body physics-grounded image-to-video generation. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer vision–ECCV 2024. Cham: Springer Nature Switzerland; 2025. p. 360–78. doi:10.1007/978-3-031-73007-8_21.
197. Gu Z, Yan R, Lu J, Li P, Dou Z, Si C, et al. Diffusion as shader: 3D-aware video diffusion for versatile video generation control. arXiv:2501.03847. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.03847.

198. Xie R, Liu Y, Zhou P, Zhao C, Zhou J, Zhang K, et al. STAR: spatial-temporal augmentation with text-to-video models for real-world video super-resolution. arXiv:2501.02976. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.02976.
199. Chen S, Guo H, Zhu S, Zhang F, Huang Z, Feng J, et al. Video depth anything: consistent depth estimation for super-long videos. arXiv:2501.12375. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.12375.
200. Yatim D, Fridman R, Bar-Tal O, Dekel T. DynVFX: augmenting real videos with dynamic content. arXiv:2502.03621. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2502.03621.
201. Zhang Y, Zhou X, Zeng Y, Xu H, Li H, Zuo W. FramePainter: endowing interactive image editing with video diffusion priors. arXiv:2501.08225. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.08225.
202. Wang R, Xu S, He T, Chen Y, Zhu W, Song D, et al. DynamicFace: high-quality and consistent video face swapping using composable 3D facial priors. arXiv:2501.08553. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.08553.
203. Bi X, Lu J, Liu B, Cun X, Zhang Y, Li W, et al. CustomTTT: motion and appearance customized video generation via test-time training. arXiv:2412.15646. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.15646.
204. Pandey K, Gadelha M, Hold-Geoffroy Y, Singh K, Mitra NJ, Guerrero P. Motion modes: what could happen next? arXiv:2412.00148. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.00148.
205. Li X, Jia X, Wang Q, Diao H, Ge M, Li P, et al. MoTrans: customized motion transfer with text-driven video diffusion models. In: Proceedings of the 32nd ACM International Conference on Multimedia (MM '24); 2024 Oct 28–Nov 1; New York, NY, USA: Association for Computing Machinery; 2024. p. 3421–30. doi:10.1145/3664647.3680718.
206. Yesiltepe H, Meral THS, Dunlop C, Yanardag P. MotionShop: zero-shot motion transfer in video diffusion models with mixture of score guidance. arXiv:2412.05355. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.05355.
207. Pondaven A, Siarohin A, Tulyakov S, Torr P, Pizzati F. Video motion transfer with diffusion transformers. arXiv:2412.07776. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.07776.
208. Wang Z, Lan Y, Zhou S, Loy CC. ObjCtrl-2.5D: training-free object control with camera poses. arXiv:2412.07721. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.07721.
209. Zhou Z, An J, Luo J. Latent-Reframe: enabling camera control for video diffusion model without training. arXiv:2412.06029. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.06029.
210. Wu J, Li X, Zeng Y, Zhang J, Zhou Q, Li Y, et al. MotionBooth: motion-aware customized text-to-video generation. arXiv:2406.17758. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2406.17758.
211. Reda F, Kontkanen J, Tabellion E, Sun D, Pantofaru C, Curless B. FILM: frame interpolation for large motion. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, editors. Computer vision–ECCV 2022. Cham: Springer Nature Switzerland; 2022. p. 250–66. doi:10.1007/978-3-031-20071-7_15.
212. Jain S, Watson D, Tabellion E, Ho?ynski A, Poole B, Kontkanen J. Video interpolation with diffusion models. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 7341–51.
213. Ma H, Mahdizadehaghdam S, Wu B, Fan Z, Gu Y, Zhao W, et al. MaskINT: video editing via interpolative non-autoregressive masked transformers. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 7403–12.
214. Feng H, Ding Z, Xia Z, Niklaus S, Abrevaya V, Black MJ, et al. Explorative inbetweening of time and space. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer vision–ECCV 2024. Cham: Springer Nature Switzerland; 2025. p. 378–95. doi:10.1007/978-3-031-73229-4_22.
215. Xing J, Liu H, Xia M, Zhang Y, Wang X, Shan Y, et al. ToonCrafter: generative cartoon interpolation. ACM Trans Graph. 2024;43(6):245:1–11. doi:10.1145/3687761.
216. Li X, Xue H, Ren P, Bo L. DiffuEraser: a diffusion model for video inpainting. arXiv:2501.10018. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.10018.
217. Wang Z, Chen X, Xu C, Zhu J, Hu X, Zhang J, et al. SVFR: a unified framework for generalized video face restoration. arXiv:2501.01235. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.01235.

218. Yang P, Zhou S, Zhao J, Tao Q, Loy CC. MatAnyone: stable video matting with consistent memory propagation. arXiv:2501.14677. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.14677.
219. Wang J, Lin Z, Wei M, Zhao Y, Yang C, Xiao F, et al. SeedVR: seeding infinity in diffusion transformer towards generic video restoration. arXiv:2501.01320. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.01320.
220. Bian Y, Zhang Z, Ju X, Cao M, Xie L, Shan Y, et al. VideoPainter: any-length video inpainting and editing with plug-and-play context control. arXiv:2503.05639. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2503.05639.
221. Teodoro S, Gunawan A, Kim SY, Oh J, Kim M. MIVE: new design and benchmark for multi-instance video editing. arXiv:2412.12877. 2024 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2412.12877.
222. Ji S, Xu W, Yang M, Yu K. 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell. 2013;35(1):221–31. doi:10.1109/tpami.2012.59.
223. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. arXiv:1711.07971. 2018 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/1711.07971.
224. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, et al. Video swin transformer. arXiv:2106.13230. 2021 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2106.13230.
225. Fei H, Wu S, Zhang H, Chua TS, Yan S. Vitron: a unified pixel-level vision LLM for understanding, generating, segmenting, editing. arXiv:2412.19806. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.19806.
226. Yuan H, Li X, Zhang T, Huang Z, Xu S, Ji S, et al. Sa2VA: marrying SAM2 with LLaVA for dense grounded understanding of images and videos. arXiv:2501.04001. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.04001.
227. Li K, Li X, Wang Y, He Y, Wang Y, Wang L, et al. VideoMamba: state space model for efficient video understanding. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer vision–ECCV 2024. Cham: Springer Nature Switzerland; 2025. p. 237–55. doi:10.1007/978-3-031-73347-5_14.
228. Ge Y, Li Y, Ge Y, Shan Y. Divot: diffusion powers video tokenizer for comprehension and generation. arXiv:2412.04432. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.04432.
229. Unterthiner T, van Steenkiste S, Kurach K, Marinier R, Michalski M, Gelly S. Towards accurate generative models of video: a new metric & challenges. arXiv:1812.01717. 2019 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/1812.01717.
230. Fischer P, Dosovitskiy A, Ilg E, Häusser P, Hazırbaş C, Golkov V, et al. FlowNet: learning optical flow with convolutional networks. arXiv:1504.06852. 2015 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/1504.06852.
231. Lubin J. A visual discrimination model for imaging system design and evaluation. In: Vision models for target detection and recognition. Series on Information Display; Volume 2, p. 245–83. World Scientific; 1995 [cited 2025 Apr 12]. Available from: https://worldscientific.com/doi/abs/10.1142/9789812831200_0010#.
232. Hessel J, Holtzman A, Forbes M, Bras RL, Choi Y. CLIPScore: a reference-free evaluation metric for image captioning. arXiv:2104.08718. 2022 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2104.08718.
233. Streijl RC, Winkler S, Hands DS. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. Multimed Syst. 2016;22(2):213–27. doi:10.1007/s00530-014-0446-1.
234. Blattmann A, Rombach R, Ling H, Dockhorn T, Kim SW, Fidler S, et al. Align your latents: high-resolution video synthesis with latent diffusion models. arXiv:2304.08818. 2023 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2304.08818.
235. Wang Z, Yuan Z, Wang X, Li Y, Chen T, Xia M, et al. MotionCtrl: a unified and flexible motion controller for video generation. In: ACM SIGGRAPH 2024 Conference Papers (SIGGRAPH '24); 2024 Jul 27–Aug 1; New York, NY, USA: Association for Computing Machinery; 2024. p. 1–11. doi:10.1145/3641519.3657518.
236. Aberman K, Weng Y, Lischinski D, Cohen-Or D, Chen B. Unpaired motion style transfer from video to animation. ACM Trans Graph. 2020;39(4):137. doi:10.1145/3386569.3392469.

237. Guizilini V, Irshad MZ, Chen D, Shakhnarovich G, Ambrus R. Zero-shot novel view and depth synthesis with multi-view geometric diffusion. arXiv:2501.18804. 2025 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2501.18804.
238. Gan Y, Yang Z, Yue X, Sun L, Yang Y. Efficient emotional adaptation for audio-driven talking-head generation. arXiv:2309.04946. 2023 [cited 2025 Apr 12]. Available from: http://arxiv.org/abs/2309.04946.
239. Google DeepMind. 2025. Music AI Sandbox, now with new features and broader access. [cited 2025 Jul 21]. Available from: https://deepmind.google/discover/blog/music-ai-sandbox-now-with-new-features-and-broader-access/.
240. Evans Z, Parker JD, Carr CJ, Zukowski Z, Taylor J, Pons J. Stable audio open. arXiv:2407.14358. 2024 [cited 2025 Jul 21]. Available from: http://arxiv.org/abs/2407.14358.
241. OpenAI Research | Milestone 2025 [cited 2025 Jul 21]. Available from: https://openai.com/research/index/milestone/.
242. Suno AI: Suno AI V4 Is Hear [The Latest Version Unveiled]. [cited 2025 Aug 17]. Available from: https://sunnoai.com/v4/.
243. Google DeepMind. 2024. Generating audio for video. [cited 2025 Jul 21]. Available from: https://deepmind.google/discover/blog/generating-audio-for-video/.
244. Copet J, Kreuk F, Gat I, Remez T, Kant D, Synnaeve G, et al. Simple and controllable music generation. arXiv:2306.05284. 2024 [cited 2025 Jul 21]. Available from: http://arxiv.org/abs/2306.05284.
245. Gong J, Zhao S, Wang S, Xu S, Guo J. ACE-Step: a step towards music generation foundation model. arXiv:2506.00045. 2025 [cited 2025 Jul 21]. Available from: http://arxiv.org/abs/2506.00045.
246. Huang Q, Park DS, Wang T, Denk TI, Ly A, Chen N, et al. Noise2Music: text-conditioned music generation with diffusion models. arXiv:2302.03917. 2023 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2302.03917.
247. Ning Z, Chen H, Jiang Y, Hao C, Ma G, Wang S, et al. DiffRhythm: blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. arXiv:2503.01183. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2503.01183.
248. Hung CY, Majumder N, Kong Z, Mehrish A, Valle R, Catanzaro B, et al. TangoFlux: super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. arXiv:2412.21037. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.21037.
249. Bai Y, Chen H, Chen J, Chen Z, Deng Y, Dong X, et al. Seed-Music: a unified framework for high quality and controlled music generation. arXiv:2409.09214. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2409.09214.
250. Wu SL, Donahue C, Watanabe S, Bryan NJ. Music controlNet: multiple time-varying controls for music generation. IEEE/ACM Transact Audio, Speech, Lang Process. 2024;32:2692–703. doi:10.1109/taslp.2024.3399026.
251. Suvée D, Vanderperren W, Jonckers V. JAsCo: an aspect-oriented approach tailored for component based software development. In: Proceedings of the 2nd International Conference on Aspect-Oriented Software Development (AOSD '03); 2023 Mar 17–21; New York, NY, USA: Association for Computing Machinery. p. 21–9. doi:10.1145/643603.643606.
252. Yuan R, Lin H, Guo S, Zhang G, Pan J, Zang Y, et al. YuE: scaling open foundation models for long-form music generation. arXiv:2503.08638. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2503.08638.
253. Sun P, Cheng S, Li X, Ye Z, Liu H, Zhang H, et al. Both ears wide open: towards language-driven spatial audio generation. arXiv:2410.10676. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2410.10676.
254. Zhang Y, Ikemiya Y, Xia G, Murata N, Martínez-Ramírez MA, Liao WH, et al. MusicMagus: zero-shot text-to-music editing via diffusion models. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI '24); 2024 Aug 3–9; Jeju, Republic of Korea. p. 7805–13. doi:10.24963/ijcai.2024/864.
255. Koo J, Wichern G, Germain FG, Khurana S, Roux JL. SMITIN: self-monitored inference-time INtervention for generative music transformers. IEEE Open J Signal Process. 2025;6:266–75. doi:10.1109/ojsp.2025.3534686.
256. Haji-Ali M, Menapace W, Siarohin A, Skorokhodov I, Canberk A, Lee KS, et al. AV-Link: temporally-aligned diffusion features for cross-modal audio-video generation. arXiv:2412.15191. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.15191.

257. Cheng HK, Ishii M, Hayakawa A, Shibuya T, Schwing A, Mitsufuji Y. Taming multimodal joint training for high-quality video-to-audio synthesis. arXiv:2412.15322. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.15322.
258. Chen Z, Seetharaman P, Russell B, Nieto O, Bourgin D, Owens A, et al. Video-guided foley sound generation with multimodal controls. arXiv:2411.17698. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2411.17698.
259. Wang B, Zhuo L, Wang Z, Bao C, Chengjing W, Nie X, et al. Multimodal music generation with explicit bridges and retrieval augmentation. arXiv:2412.09428. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.09428.
260. Gramaccioni RF, Marinoni C, Postolache E, Comunità M, Cosmo L, Reiss JD, et al. Stable-V2A: synthesis of synchronized sound effects with temporal and semantic controls. arXiv:2412.15023. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.15023.
261. Li S, Yang B, Yin C, Sun C, Zhang Y, Dong W, et al. VidMusician: video-to-music generation with semantic-rhythmic alignment via hierarchical visual features. arXiv:2412.15023. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.06296.
262. Zhang C, Wang C, Zhang J, Xu H, Song G, Xie Y, et al. DREAM-Talk: diffusion-based realistic emotional audio-driven method for single image talking face generation. arXiv:2312.13578. 2023 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2312.13578.
263. Cheng X, Wang X, Wu Y, Wang Y, Song R. LoVA: long-form video-to-audio generation. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2025 Apr 6–11; Hyderabad, India: ICASSP. p. 1–5.
264. Comanducci L, Bestagini P, Tubaro S. FakeMusicCaps: a dataset for detection and attribution of synthetic music generated via text-to-music models. arXiv:2409.10684. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2409.10684.
265. Ciranni R, Mariani G, Mancusi M, Postolache E, Fabbro G, Rodolà E, et al. COCOLA: coherence-oriented contrastive learning of musical audio representations. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2025 Apr 6–11; Hyderabad, India. p. 1–5.
266. Ni J, Song Y, Ghosal D, Li B, Zhang DJ, Yue X, et al. MixEval-X: any-to-any evaluations from real-world data mixtures. arXiv:2410.13754. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2410.13754.
267. Ji S, Yang X, Luo J. A survey on deep learning for symbolic music generation: representations, algorithms, evaluations, and challenges. ACM Comput Surv. 2023;56(1):7:1–39. doi:10.1145/3597493.
268. Yang LC, Lerch A. On the evaluation of generative models in music. Neural Comput & Applic. 2020;32(9):4773–84.
269. Yin Z. New evaluation methods for automatic music generation [Ph.D thesis]. University of York; 2022 [cited 2025 Apr 13]. Available from: https://etheses.whiterose.ac.uk/id/eprint/31507/.
270. Liu X, Chen Q, Wu X, Liu Y, Liu Y. CNN based music emotion classification. arXiv:1704.05665. 2017 [cited 2025 Apr 13]. Available from: http://arxiv.org/abs/1704.05665.
271. Yading S, Simon D, Marcus P. Evaluation of musical features for emotion classification. In: 13th International Society for Music Information Retrieval Conference; 2012 Oct 8–12; Porto, Portugal. p. 523–8.
272. Hawthorne C, Simon I, Roberts A, Zeghidour N, Gardner J, Manilow E, et al. Multi-instrument music synthesis with spectrogram diffusion. arXiv:2206.05408. 2022 [cited 2025 Apr 13]. Available from: http://arxiv.org/abs/2206.05408.
273. Liu X. Application of audio signal processing technology in music synthesis. J Elect Syst. 2024;20(9s):648–54.
274. Valevski D, Leviathan Y, Arar M, Fruchter S. Diffusion models are real-time game engines. arXiv:2408.14837. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2408.14837.
275. Oasis: A Universe in a Transformer. Oasis: a universe in a transformer. [cited 2025 Apr 8]. Available from: https://www.decart.ai/articles/oasis-interactive-ai-video-game-model.
276. Yang M, Li J, Fang Z, Chen S, Yu Y, Fu Q, et al. Playable game generation. arXiv:2412.00887. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.00887.
277. Google DeepMind. 2025. Genie 2: A large-scale foundation world model. [cited 2025 Apr 8]. Available from: https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/.

278. Liang H, Cao J, Goel V, Qian G, Korolev S, Terzopoulos D, et al. Wonderland: navigating 3D scenes from a single image. arXiv:2412.12091. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.12091.
279. Sun W, Chen S, Liu F, Chen Z, Duan Y, Zhang J, et al. DimensionX: create any 3D and 4D scenes from a single image with controllable video diffusion. arXiv:2411.04928. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2411.04928.
280. Deng B, Tucker R, Li Z, Guibas L, Snavely N, Wetzstein G. Streetscapes: large-scale consistent street view generation using autoregressive video diffusion. In: ACM SIGGRAPH 2024 Conference Papers; 2024 Jul 27–Aug 1; New York, NY, USA: Association for Computing Machinery. p. 1–11. (SIGGRAPH '24) doi:10.1145/3641519.3657513.
281. Gupta V, Man Y, Wang YX. PaintScene4D: consistent 4D scene generation from text prompts. arXiv:2412.04471. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.04471.
282. Bai J, Xia M, Wang X, Yuan Z, Fu X, Liu Z, et al. SynCamMaster: synchronizing multi-camera video generation from diverse viewpoints. arXiv:2412.07760. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.07760.
283. Lin J, Yang X, Chen M, Xu Y, Yan D, Wu L, et al. Kiss3DGen: repurposing image diffusion models for 3D asset generation. arXiv:2503.01370. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2503.01370.
284. Zhou Y, Bu J, Ling P, Zhang P, Wu T, Huang Q, et al. Light-a-video: training-free video relighting via progressive light fusion. arXiv:2502.08590. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2502.08590.
285. Wu CH, Chen YJ, Chen YH, Lee JY, Ke BH, Mu CWT, et al. AuraFusion360: augmented unseen region alignment for reference-based 360° unbounded scene inpainting. arXiv:2502.05176. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2502.05176.
286. Condor J, Speierer S, Bode L, Bozic A, Green S, Didyk P, et al. Don't splat your gaussians: volumetric ray-traced primitives for modeling and rendering scattering and emissive media. ACM Trans Graph. 2025;44(1):10:1–10:17. doi:10.1145/3711853.
287. Yang J, Sax A, Liang KJ, Henaff M, Tang H, Cao A, et al. Fast3R: towards 3D reconstruction of 1000+ images in one forward pass. arXiv:2501.13928. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.13928.
288. Cong W, Zhu H, Wang K, Lei J, Stearns C, Cai Y, et al. VideoLifter: lifting videos to 3D with fast hierarchical stereo alignment. arXiv:2501.01949. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.01949.
289. Lyu W, Li X, Kundu A, Tsai YH, Yang MH. Gaga: group any gaussians via 3D-aware memory bank. arXiv:2404.07977. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2404.07977.
290. Qiu L, Zhu S, Zuo Q, Gu X, Dong Y, Zhang J, et al. AniGS: animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. arXiv:2412.02684. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.02684.
291. Kant Y, Weber E, Kim JK, Khirodkar R, Zhaoen S, Martinez J, et al. Pippo: high-resolution multi-view humans from a single image. arXiv:2502.07785. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2502.07785.
292. Cha H, Lee I, Joo H. PERSE: personalized 3D generative avatars from a single portrait. arXiv:2412.21206. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.21206.
293. Ren J, Sundaresan P, Sadigh D, Choudhury S, Bohg J. Motion tracks: a unified representation for human-robot transfer in few-shot imitation learning. arXiv:2501.06994. 2025 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2501.06994.
294. Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative agents: interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23); 2023 Oct 29–Nov 1; New York, NY, USA: Association for Computing Machinery. p. 1–22. doi:10.1145/3586183.3606763.
295. Google DeepMind. 2025. A generalist AI agent for 3D virtual environments. [cited 2025 Apr 8]. Available from: https://deepmind.google/discover/blog/sima-generalist-ai-agent-for-3d-virtual-environments/.
296. Zheng J, Wang J, Zhao F, Zhang X, Liang X. Dynamic try-on: taming video virtual try-on with dynamic attention mechanism. arXiv:2412.09822. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.09822.
297. Wang L, Zheng W, Du D, Zhang Y, Ren Y, Jiang H, et al. Stag-1: towards realistic 4D driving simulation with video generation model. arXiv:2412.05280. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.05280.

298. GiiNEX. [cited 2025 Apr 13]. Available from: https://giinex.tencent.com/#/index.
299. Kanervisto A, Bignell D, Wen LY, Grayson M, Georgescu R, Valcarcel Macua S, et al. World and human action models towards gameplay ideation. Nature. 2025;638(8051):656–63. doi:10.1038/s41586-025-08600-3.
300. Layer | Game Art Without Limits. [cited 2025 Apr 13]. Available from: https://layer.ai/.
301. Scenario - Take complete control of your AI workflows. [cited 2025 Apr 13]. Available from: https://www.scenario.com/.
302. Gemini. Gemini Apps' release updates & improvements. [cited 2025 Aug 17]. Available from: https://gemini.google/release-notes/.
303. Claude 3.7 Sonnet and Claude Code. [cited 2025 Jul 21]. Available from: https://www.anthropic.com/news/claude-3-7-sonnet.
304. DeepSeek-AI, Guo D, Yang D, Zhang H, Song J, Zhang R, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. arXiv:2501.12948. 2025 [cited 2025 Apr 13]. Available from: http://arxiv.org/abs/2501.12948.
305. mistralai/Mistral-Large-Instruct-2407. Hugging Face. [cited 2025 Jul 21]. Available from: https://huggingface.co/mistralai/Mistral-Large-Instruct-2407.
306. gpt-4-5-system-card-2272025.pdf. [cited 2025 Jul 21]. Available from: https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf.
307. News | xAI. [cited 2025 Jul 21]. Available from: https://x.ai/news.
308. Agrawal P, Antoniak S, Hanna EB, Bout B, Chaplot D, Chudnovsky J, et al. Pixtral 12B. arXiv:2410.07073. 2024 [cited 2025 Jul 21]. Available from: http://arxiv.org/abs/2410.07073.
309. meta-llama/llama3. Meta Llama; 2025 [cited 2025 Jul 21]. Available from: https://github.com/meta-llama/llama3.
310. Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, et al. Mixtral of experts. arXiv:2401.04088. 2024 [cited 2025 Jul 21]. Available from: http://arxiv.org/abs/2401.04088.
311. License and copyright - arXiv info. [cited 2025 Jul 21]. Available from: https://info.arxiv.org/help/license/index.html#licenses-available.
312. Team G, Kamath A, Ferret J, Pathak S, Vieillard N, Merhej R, et al. Gemma 3 technical report. arXiv:2503.19786. 2025 [cited 2025 Jul 21]. Available from: http://arxiv.org/abs/2503.19786.
313. Abdin M, Aneja J, Behl H, Bubeck S, Eldan R, Gunasekar S, et al. Phi-4 technical report. arXiv:2412.08905. 2024 [cited 2025 Jul 21]. Available from: http://arxiv.org/abs/2412.08905.
314. Databricks. 2024 introducing DBRX: a new state-of-the-art open LLM. [cited 2025 Jul 21]. Available from: https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm.
315. Nvidia, Adler B, Agarwal N, Aithal A, Anh DH, Bhattacharya P, et al. Nemotron-4 340B technical report. arXiv:2406.11704. 2024 [cited 2025 Jul 21]. Available from: http://arxiv.org/abs/2406.11704.
316. Team TH, Liu A, Zhou B, Xu C, Zhou C, Zhang C, et al. Hunyuan-TurboS: advancing large language models through mamba-transformer synergy and adaptive chain-of-thought. arXiv:2505.15431. 2025 [cited 2025 Jul 21]. Available from: http://arxiv.org/abs/2505.15431.
317. llm.hunyuan.T1. llm.hunyuan.T1. [cited 2025 Jul 21]. Available from: https://tencent.github.io/llm.hunyuan.T1/README_EN.html.
318. GitHub. Build software better, together. [cited 2025 Jul 21]. Available from: https://github.com.
319. Kimi K2: open agentic intelligence. [cited 2025 Jul 21]. Available from: https://moonshotai.github.io/Kimi-K2/.
320. AlibabaCloud. Tongyi Qianwen (Qwen) - Alibaba Cloud. [cited 2025 Aug 17]. Available from: https://www.alibabacloud.com/solutions/generative-ai/qwen.
321. Yang A, Yang B, Hui B, Zheng B, Yu B, Zhou C, et al. Qwen2 technical report. arXiv:2407.10671. 2024 [cited 2025 Jul 21]. Available from: http://arxiv.org/abs/2407.10671.
322. MiniMax, Li A, Gong B, Yang B, Shan B, Liu C, et al. MiniMax-01: scaling foundation models with lightning attention. arXiv:2501.08313. 2025 [cited 2025 Jul 21]. Available from: http://arxiv.org/abs/2501.08313.
323. Zhao Z, Song S, Duah B, Macbeth J, Carter S, Van MP, et al. More human than human: LLM-generated narratives outperform human-LLM interleaved narratives. In: Proceedings of the 15th Conference on Creativity

and Cognition; 2023 Jun 19–21; New York, NY, USA: Association for Computing Machinery. p. 368–70. doi:10.1145/3591196.3596612.
324. Google. 2024. Introducing Gemini 2.0: our new AI model for the agentic era. [cited 2025 Apr 13]. Available from: https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.
325. Anand R. Unveiling google's gemini 2.0: a comprehensive study of its multimodal AI design, advanced architecture, and real-world applications. [cited 2025 Apr 13]. Available from: https://www.researchgate.net/publication/387089907_Unveiling_Google's_Gemini_20_A_Comprehensive_Study_of_its_Multimodal_AI_Design_Advanced_Architecture_and_Real-World_Applications.
326. Introducing Claude 3.5 Sonnet. [cited 2025 Apr 13]. Available from: https://www.anthropic.com/news/claude-3-5-sonnet.
327. Virtual YouTuber Wiki. 2025. Neuro-sama. [cited 2025 Apr 13]. Available from: https://virtualyoutuber.fandom.com/wiki/Neuro-sama.
328. Magic-Emerge/luna-ai. Magic Emerge. 2025 [cited 2025 Apr 13]. Available from: https://github.com/Magic-Emerge/luna-ai.
329. Desai AP, Ravi T, Luqman M, Sharma M, Kota N, Yadav P. Gen-AI for user safety: a survey. In: 2024 IEEE International Conference on Big Data (BigData); 2024 Dec 15–18; Washington, DC, USA. p. 5315–24.
330. Karagoz A. Ethics and technical aspects of generative AI models in digital content creation. arXiv:2412.16389. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.16389.
331. Villate-Castillo G, Ser JD, Sanz B. A collaborative content moderation framework for toxicity detection based on conformalized estimates of annotation disagreement. arXiv:2411.04090. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2411.04090.
332. Banchio PR. Legal, ethical and practical challenges of AI-driven content moderation. Rochester, NY: Social Science Research Network. 2024 [cited 2025 Apr 8]. Available from: https://papers.ssrn.com/abstract=4984756.
333. Lloyd T, Reagle J, Naaman M. "There has to be a lot that we're missing": moderating AI-generated content on reddit. arXiv:2311.12702. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2311.12702.
334. Abdali S, Anarfi R, Barberan C, He J. Decoding the AI pen: techniques and challenges in detecting AI-generated text. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24); 2024 Aug 25–29; New York, NY, USA: Association for Computing Machinery. p. 6428–36. doi:10.1145/3637528.3671463.
335. Hao S, Han W, Jiang T, Li Y, Wu H, Zhong C, et al. Synthetic data in AI: challenges, applications, and ethical implications. arXiv:2401.01629. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2401.01629.
336. Appel RE. Generative AI regulation can learn from social media regulation. arXiv:2412.11335. 2024 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/2412.11335.
337. Hagerty A, Rubinov I. Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. arXiv:1907.07892. 2019 [cited 2025 Apr 8]. Available from: http://arxiv.org/abs/1907.07892.
338. Arora S, Arora S, Hastings J. The psychological impacts of algorithmic and AI-driven social media on teenagers: a call to action. In: 2024 IEEE Digital Platforms and Societal Harms (DPSH); 2024 Oct 14–15; Washington, DC, USA. p. 1–7. doi:10.1109/dpsh60098.2024.10774922.
339. Khlaif ZN. Ethical concerns about using AI-generated text in scientific research. Rochester, NY: Social Science Research Network; 2023 [cited 2025 Apr 8]. Available from: https://papers.ssrn.com/abstract=4387984.
340. Illia L, Colleoni E, Zyglidopoulos S. Ethical implications of text generation in the age of artificial intelligence. Business Ethics Environ Respons. 2023;32(1):201–10.
341. Gillespie T. Content moderation, AI, and the question of scale. Big Data & Society [Internet]. 2020 Aug 21 [cited 2025 Apr 8];7(2). Available from: https://journals.sagepub.com/doi/full/10.1177/2053951720943234.
342. Stout DW. How generative AI Has transformed creative work: a comprehensive study. Magai. 2025 [cited 2025 May 20]. Available from: https://magai.co/generative-ai-has-transformed-creative-work/.
343. 50 arguments against the use of AI in creative fields. [cited 2025 May 20]. Available from: https://aokistudio.com/50-arguments-against-the-use-of-ai-in-creative-fields.html.

344. Thambaiya N, Kariyawasam K, Talagala C. Copyright law in the age of AI: analysing the AI-generated works and copyright challenges in Australia. Int Rev Law Comput Technol. 2025;35(2):1–26. doi:10.1080/13600869.2025.2486893.
345. Topics | European Parliament. 2023. EU AI Act: first regulation on artificial intelligence. [cited 2025 Jun 4]. Available from: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.
346. GOV.UK. Artificial Intelligence and Intellectual Property: copyright and patents. [cited 2025 Jun 4]. Available from: https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/artificial-intelligence-and-intellectual-property-copyright-and-patents.
347. kimjones. Adapting models to handle cultural variations in language and context. Welocalize. 2024 [cited 2025 May 20]. Available from: https://www.welocalize.com/insights/adapting-models-to-handle-cultural-variations-in-language-and-context/.
348. Trinity College Dublin. Generative AI models are encoding biases and negative stereotypes in their users. [cited 2025 May 20]. Available from: https://www.tcd.ie/news_events/articles/2023/generative-ai-models-are-encoding-biases-and-negative-stereotypes-in-their-users/.
349. Spotify. Long Reads. [cited 2025 Jun 4]. Available from: https://open.spotify.com/show/4mqFZZV8DMVeN80VwUumP4.
350. Studio Ghibli AI Art Sparks Big Debate Over Creativity and Copyright. Studio ghibli AI art sparks big debate over creativity and copyright. [cited 2025 May 20]. Available from: https://opentools.ai/news/studio-ghibli-ai-art-sparks-big-debate-over-creativity-and-copyright.
351. Castro D. Critics of generative AI are worrying about the wrong IP issues. 2023 Mar [cited 2025 May 20]. Available from: https://itif.org/publications/2023/03/20/critics-of-generative-ai-are-worrying-about-the-wrong-ip-issues/.
352. Rice AB. Miller library libguides: AI as a research tool: ethical and legal considerations. [cited 2025 Jun 4]. Available from: https://washcoll.libguides.com/c.php?g=1414339&p=10478589.
353. How animation industry can be transformed by generative AI. [cited 2025 May 20]. Available from: https://www.e2enetworks.com/blog/how-animation-industry-can-be-transformed-by-generative-ai.
354. AI Takes Center Stage in Animation's Future: Top Tools of 2024 Revealed!. AI Takes Center Stage in Animation's Future: Top Tools of 2024 Revealed!. [cited 2025 May 20]. Available from: https://opentools.ai/news/ai-takes-center-stage-in-animations-future-top-tools-of-2024-revealed.
355. Team SP. SoA survey reveals a third of translators and quarter of illustrators losing work to AI. The Society of Authors; 2024 [cited 2025 May 20]. Available from: https://societyofauthors.org/2024/04/11/soa-survey-reveals-a-third-of-translators-and-quarter-of-illustrators-losing-work-to-ai/.
356. Hall J, Schofield D. The value of creativity: human produced art vs. AI-generated art. Art Design Rev. 2024;13(1):65–88. doi:10.4236/adr.2025.131005.
357. Diffey H. ScreenRant. 2025 [cited 2025 Jun 4]. Amid Toei's AI Controversy, the Anime Industry Is Pushing Back: "Aren't We Shooting Ourselves In the Foot?". Available from: https://screenrant.com/anime-ai-japan-backlash-toei-animation-industry-future/.
358. Economic potential of generative AI | McKinsey. [cited 2025 May 21]. Available from: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier.
359. AI Development Cost: A Comprehensive Overview for 2025. Prismetric; 2025 [cited 2025 May 21]. Available from: https://www.prismetric.com/ai-development-cost/.
360. Outshift by Cisco. Outshift | AI infrastructure: Prepare your organization for transformation. [cited 2025 May 21]. Available from: https://outshift.com/blog/ai-infrastructure-how-to-prepare-your-organization-for-transformation.
361. Animation World Network. The future unscripted: the impact of generative artificial intelligence on entertainment industry jobs. [cited 2025 May 21]. Available from: https://www.awn.com/tag/future-unscripted-impact-generative-artificial-intelligence-entertainment-industry-jobs.

362. Wolff E. Educating artists about AI in animation. Animation Magazine. 2025. [cited 2025 May 21]. Available from: https://www.animationmagazine.net/2025/03/educating-artists-about-ai-in-animation/.
363. Fine-Arts-2022.pdf. [cited 2025 Jun 5]. Available from: https://coloradostatefair.com/wp-content/uploads/2022/05/Fine-Arts-2022.pdf.
364. Christies. Christies.com. JESSE WOOLSTON. [cited 2025 Jun 5]. Available from: https://onlineonly.christies.com/s/augmented-intelligence/jesse-woolston-22/250105.
365. Christies. Christies.com. HOLLY HERNDON (B. 1980) AND MAT DRYHURST (B. 1984). [cited 2025 Jun 5]. Available from: https://onlineonly.christies.com/s/augmented-intelligence/holly-herndon-b-1980-mat-dryhurst-b-1984-3/249745.
366. pixiv. Popular illustrations and manga tagged "AI". [cited 2025 Aug 17]. Available from: https://www.pixiv.net/en/tags/AI.
367. ArtStation. ArtStation-Explore. [cited 2025 Jun 5]. Available from: https://artstation.com/.
368. Anime News Network. 2025. Novelist otsuichi co-directs generaidoscope, omnibus film produced entirely with generative AI. [cited 2025 Jun 5]. Available from: https://www.animenewsnetwork.com/news/2024-07-13/novelist-otsuichi-co-directs-generaidoscope-omnibus-film-produced-entirely-with-generative-ai/.213069.
369. *Twins Hinahima*. In: Wikipedia; 2025 [cited 2025 Jun 5]. Available from: https://en.wikipedia.org/w/index.php?title=Twins_Hinahima&oldid=1284878942.
370. Deep Learning Super Sampling. In: Wikipedia. 2025 [cited 2025 Jun 5]. Available from: https://en.wikipedia.org/w/index.php?title=Deep_Learning_Super_Sampling&oldid=1291331644.
371. *Fossilized Wonders*. In: Wikipedia. 2025. [cited 2025 Jun 5]. Available from: https://en.wikipedia.org/w/index.php?title=Fossilized_Wonders&oldid=1291840646.