Available Online: 26 January 2026 DOI: 10.54254/2977-3903/2026.31448

A survey on text-driven visual generation: advances, frameworks, and future directions

Haoting Jv

College of Computer and Information Science/College of Software, Southwest University, Chongqing, China

jht18963139708@email.swu.edu.cn

Abstract. After Denoising Diffusion Probabilistic Models (DDPM) outperformed Generative Adversarial Networks (GANs), diffusion models have evolved into the backbone of text-guided visual generation, with Stable Diffusion and DALL·E 2 alleviating key technical constraints. Despite remarkable advances in Text-to-Image (T2I) and Text-to-Video (T2V) tasks, critical gaps remain unaddressed. This paper conducts a systematic review of diffusion-based T2I and T2V technologies, synthesises the latest advances in related technologies, and proposes a "Technical Module-Application-Evaluation" framework to link technical breakthroughs with real-world applications. It also highlights under-researched fields and corresponding evaluation benchmarks, offering an integrated technical landscape to guide the equitable and reliable industrialisation of text-driven visual generation technologies.

Keywords: diffusion models, text-to-image generation, text-to-video generation, controllable visual generation, physics-aware generation

1. Introduction

Diffusion models have reshaped text-driven visual generation by addressing the inherent drawbacks of GANs, driven by two pivotal innovations: latent-space computation in Stable Diffusion reduces resource demands, while Contrastive Language-Image Pre-training (CLIP) - Large Language Model (LLM) integration in DALL·E 2 enhances text-image alignment. These models now underpin applications across diverse industrial sectors, enabling high-resolution production, mobile deployment, and clinical-grade medical visual synthesis.

However, major challenges persist: the quality-efficiency trade-off in practical deployment, the under-representation of non-Western cultures in mainstream models, limited physical realism in generated content, and outdated academic surveys that overlook 2025-era progress and fail to connect technological advances with real-world requirements. To address these issues, this paper presents a comprehensive and up-to-date review of diffusion-based T2I and T2V technologies.

This work makes three core contributions: it is the first to systematically integrate post-2025 advances in text-driven visual generation, covering key innovations such as linear attention mechanisms, weak-to-strong training, and multimodal collaborative control to fill gaps left by earlier surveys; it introduces the "Technical Module-Application-Evaluation" framework, bridging the disconnect between technical innovation, application scenarios, and fragmented evaluation practices to align development pathways with practical needs; it also foregrounds long-neglected areas, including non-Western cultural representation and physical plausibility, establishing targeted evaluation benchmarks and optimisation routes to guide equitable, reliable, and practical industrialisation of related technologies.

2. Overview

2.1. Development foundation of diffusion models

Diffusion models have become the mainstream approach for text-driven visual generation after surpassing GANs, with T2I and T2V technologies evolving along parallel trajectories. For T2I, the shift from pixel-space to latent-space computation has significantly improved resource efficiency, while the integration of vision-language models has further enhanced text-image alignment and output diversity. Architectural innovations and optimised training paradigms have improved these models' ability to interpret complex text prompts, synthesise ultra-high-resolution content, and scale across both data centre and consumer-grade hardware.

For T2V, research has focused on extending diffusion mechanisms to temporal data, with inter-frame coherence recognised as a key technical challenge. Early progress enabled the generation of multi-second video clips, while subsequent breakthroughs introduced specialised architectures and compression techniques that upgraded output quality and scalability. Key milestones include the development of physically realistic "world simulation," real-time generation capabilities, and mobile-end deployment - advances that have shifted T2V from laboratory experimentation to industrial application by addressing bottlenecks in temporal length, output quality, and computational efficiency.

2.2. Core technical logic and application context

T2I's technical advances cluster around three interconnected directions. Efficient high-resolution generation is central, balancing output quality with computational cost. Multilingual text rendering is another key focus, addressing typographic and semantic challenges across languages. Interactive controllability completes the set, enhancing users' control over spatial and stylistic elements. These advances have been widely applied: they mitigate data shortages in healthcare, simplify animation content creation, and reduce biases that hinder cross-cultural communication.

T2V's development revolves around three core objectives. Long-video generation is a primary pursuit, maintaining temporal coherence across extended footage. Controllable synthesis is equally prioritised, enabling precise control of camera movements and scene elements. Physics-aware generation forms the third pillar, integrating adherence to physical laws. These innovations have been applied across sectors: accelerating gaming prototyping, providing safe simulations for medical training, and supporting industrial simulation for production optimisation and fault prediction. Figure 1 provides an overview of the survey on text-driven visual generation.

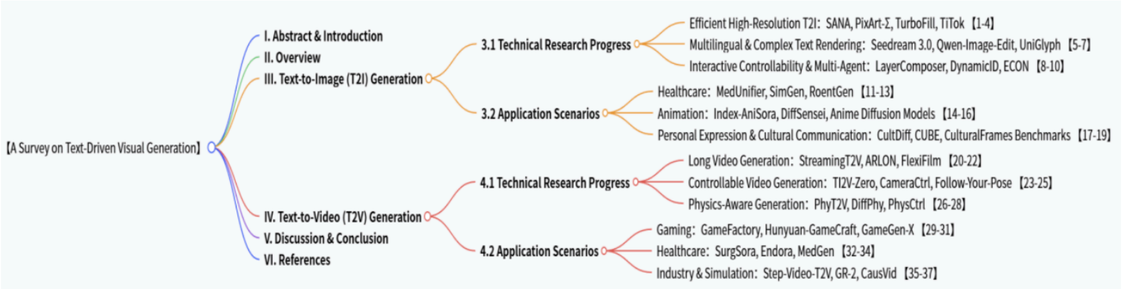
【A Survey on Text-Driven Visual Generation】
I. Abstract & Introduction
II. Overview
III. Text-to-Image (T2I) Generation
IV. Text-to-Video (T2V) Generation
V. Discussion & Conclusion
VI. References
3.1 Technical Research Progress
3.2 Application Scenarios
4.1 Technical Research Progress
4.2 Application Scenarios
Efficient High-Resolution T2I: SANA, PixArt-Σ, TurboFill, TiTok 【1-4】
Multilingual & Complex Text Rendering: Seedream 3.0, Qwen-Image-Edit, UniGlyph 【5-7】
Interactive Controllability & Multi-Agent: LayerComposer, DynamicID, ECON 【8-10】
Healthcare: MedUnifier, SimGen, RoentGen 【11-13】
Animation: Index-AniSora, DiffSensei, Anime Diffusion Models 【14-16】
Personal Expression & Cultural Communication: CultDiff, CUBE, CulturalFrames Benchmarks 【17-19】
Long Video Generation: StreamingT2V, ARLON, FlexiFilm 【20-22】
Controllable Video Generation: TI2V-Zero, CameraCtrl, Follow-Your-Pose 【23-25】
Physics-Aware Generation: PhyT2V, DiffPhy, PhysCtrl 【26-28】
Gaming: GameFactory, Hunyuan-GameCraft, GameGen-X 【29-31】
Healthcare: SurgSora, Endora, MedGen 【32-34】
Industry & Simulation: Step-Video-T2V, GR-2, CausVid 【35-37】

Figure 1. Overview of the survey on text-driven visual generation

3. Text to image generation

3.1. Research progress in technical aspects

3.1.1. Efficient and high-resolution diffusion models

With the rising demand for industrial-grade Text-to-Image (T2I) technologies, 2025-era research has focused on balancing high-resolution quality, computational efficiency, and lightweight deployment, with core breakthroughs in architecture innovation and training paradigm optimisation.

SANA [1], a linear diffusion Transformer, reduces attention complexity from O(N^2) to O(N) via linear attention. Its 32× compression AutoEncoder and lightweight Gemma text encoder cut 4K generation memory usage by 16×, enabling real-time 1024×1024 synthesis and outperforming Flux-12B with 106× faster inference and 18% lower Fréchet Inception Distance (FID) at 4K resolution.

PixArt-Σ [2], a Diffusion Transformer (DiT) model trained via a weak-to-strong paradigm based on PixArt-α, integrates 2.3 million high-quality 4K images and Share-Captioner. It adopts key-value compression to reduce training and inference time by 34%, achieving 72% and 65% win rates over Stable Diffusion XL (SDXL) and Midjourney in human preference tests with only 0.6B parameters.

TurboFill [3], an efficient inpainting method, adds a lightweight Low-Rank Adaptation (LoRA) adapter to the few-step distilled model DMD-2. Trained via 3-step adversarial learning, it is 8× faster than BrushNet at 512×512 resolution and improves structural consistency by 22% in complex occlusions.

TiTok architecture [4] compresses images into 32 discrete token sequences via a 1D Transformer. Without training an independent generative model, it matches Stable Diffusion 2.1's FID score (28.7) while being 10× faster through CLIP score optimisation.

Thanks to innovations such as linear attention and lightweight encoders, these models effectively balance high-resolution generation and computational efficiency, facilitating practical deployment across various devices. However, ultra-high-resolution synthesis still faces trade-offs with inference speed in extreme cases, and fine-grained detail preservation can lag behind heavier models.

3.1.2. Multilingual and complex text rendering

Multilingual adaptation and complex text rendering remain core challenges for T2I industrial applications, especially typography collapse in non-Latin languages such as Chinese and reduced clarity for small fonts. Recent studies have expanded large-scale use in commercial design and cross-cultural communication through innovative mechanisms.

ByteDance's Seedream 3.0 [5] introduces Cross-modal Rotational Position Embedding (Cross-modal RoPE), modelling text tokens as 2D spatial sequences to achieve 94% Chinese character usability. It adopts mixed-resolution training (256^2–2048^2) to reduce small-font (< 12pt) collapse, outperforming Midjourney V6.1 by 42% in dense layout text-image alignment and surpassing GPT-4o in Bench-377.

Alibaba's Qwen-Image-Edit [6], an International Conference on Computer Vision (ICCV) 2025 Spotlight work, uses a dual-encoding mechanism (high-level semantics + low-level reconstruction) for multilingual editing in Chinese, English, Japanese, and Korean. Its MSRoPE enhances character-background fusion by 35% in calligraphy style transfer compared with Seedream 3.0, achieving 0.946 Chinese semantic accuracy on OneIG-Bench at one-quarter the size of DALL-E 3.

UniGlyph [7], published at ICCV 2025, proposes a unified segmentation-mask conditioning framework that eliminates reliance on font annotation, supporting over 10 languages. Built on Glyph-MM-3M and Poster-100K datasets, it improves small-font (< 8pt) clarity by 37% over baselines, ranking first in ICCV's Visual Text Benchmark with 89.2 points.

Advances in cross-modal embedding techniques and dual-encoding frameworks have significantly enhanced rendering of non-Latin languages and small-font clarity, thereby driving cross-cultural commercial applications. Nonetheless, T2I performance remains inconsistent for low-resource languages, and rendering specialised terminology or mixed-style text still poses challenges for achieving accurate and aesthetically coherent results.

3.1.3. Advances in interactive controllability and multi-agent generation

Interactive controllability and multi-agent generation are key to scaling AI-driven content creation, as researchers tackle spatial misalignment, identity inconsistency, and collaboration inefficiency through novel frameworks.

LayerComposer [8] introduces a spatially-aware layered canvas and locking system, enabling intuitive manipulation of multi-subject positions, scales, and occlusions. Leveraging transparent latent pruning, it reduces memory usage to one-quarter of baselines, supports real-time editing for four or more subjects, and achieves 83.3% user approval for natural element interactions.

DynamicID [9], an ICCV 2025 Spotlight work, enables zero-shot multi-ID personalisation through Semantic-Activated Attention (SAA) and an Identity-Motion Reconfigurator (IMR). It maintains 94.6% facial feature retention on the Visual Text Benchmark and allows independent emotion adjustment without multi-ID training data.

T2I-Copilot [10] proposes a training-free multi-agent T2I system with interactive controllability, adopting a collaborative architecture of specialized agents. It surpasses FLUX1.1-pro by 6.17% at only 16.59% of the cost, matching commercial models like RecraftV3 while enabling human-in-the-loop fine-grained control.

Spatially-aware frameworks and multi-agent coordination mechanisms enhance user control over scene elements and reduce memory usage, making personalised content creation more intuitive. However, complex tasks may still encounter reasoning delays in multi-agent collaboration, and fine-grained stylistic adjustments remain less precise than manual editing.

3.2. Application of text-to-image generation

3.2.1. Healthcare

Text-to-Image (T2I) generation is reshaping healthcare by addressing data shortages, high annotation costs, and inefficient clinical communication through targeted technical improvements.

MedUnifier [11], presented at Conference on Computer Vision and Pattern Recognition (CVPR) 2025, uses a vision-and-language pre-training framework optimised for medical data with discrete visual representations. It integrates image-text contrastive alignment and vector quantisation, generating high-quality medical images that achieve 0.912 Area Under the Curve (AUC) in disease classification and 0.68 Bilingual Evaluation

Understudy-4 (BLEU-4) in report generation on Medical Information Mart for Intensive Care-Chest X-Ray (MIMIC-CXR), with 30% lower computational complexity.

SimGen [12], published in Medical Image Analysis, proposes a diffusion-based framework that generates surgical images and segmentation masks simultaneously. It leverages cross-correlation priors and CFL regularisation, improving segmentation accuracy by 15% over baseline GANs on CholecSeg8k and reducing manual annotation requirements by 90%.

TIGER [13], published in Nature Communications, is a text-guided diffusion model for rare thyroid cancer diagnosis. It adopts a dual text-image conditioning framework and two-stage training, boosting FTC detection by 12.3% and ATC by 10.7%, with 7.4% higher overall accuracy and 9.8% better specificity, while protecting privacy via synthetic data augmentation.

T2I models tackle key healthcare challenges by generating high-quality medical images and segmentation masks, easing data scarcity and supporting rare disease simulation. However, their generated content often lacks edge-case pathological details, and these models lack sufficient validation across diverse clinical populations for widespread clinical adoption.

3.2.2. Animation

To address visual incoherence, character inconsistency, and poor editability in anime creation, researchers are integrating AI into anime production with specialised frameworks and multi-modal pipelines.

LASER [14], uses an LLM-driven attention control framework for tuning-free text-conditioned image-to-animations. It incorporates four transformation types via LLM-controlled attention injection - achieving 92.3% user preference for text alignment, while boosting inference speed by 3.2× and cutting computational costs by 75%.

DiffSensei [15] connects multi-modal LLMs with diffusion models to produce customised manga. Trained on MangaZero, it maintains character consistency, optimises dialogue layout, and converts simple text into complex storyboards with strong style fidelity and text alignment.

LayerAnimate [16], uses a layered diffusion architecture for text-driven animation generation with independent layer control. It incorporates cross-layer attention fusion and LLM-powered layer assignment - achieving 21.7% higher layer independence, while cutting FVD by 15.2% and boosting production efficiency by 80% for anime creators.

These specialised frameworks ensure character consistency and style fidelity, slashing animation production cycles from weeks to seconds. However, replicating niche artistic styles remains challenging, and long sequences may still require manual adjustments to maintain visual coherence.

3.2.3. Personal expression and cultural communication

Researchers use targeted benchmarks and evaluation metrics to fix T2I models' cultural adaptability issues, tackling biases and misaligned cultural expectations.

Bayramli et al. [17] created the CultDiff benchmark covering cultural traits from 10 countries. It reveals that mainstream models underrepresent non-Western cultures, and their CultDiff-S metric boosts cross-cultural performance by 41% in culturally fine-tuned models.

Kannen et al. [18] built the first cultural competence framework for T2I models, measuring performance through cultural awareness and diversity. Their CUBE benchmark (8 countries, 3 domains) exposes stereotypes in representations of non-dominant cultures using quality-weighted Vendi scores.

Nayak et al. [19] combines 983 culture-specific prompts and over 10,000 human annotations from 10 countries. It shows that mainstream models have a 44% average failure rate in meeting cultural expectations, with 68% of failures being explicit.

Cultural benchmarks and fine-tuning methods have reduced stereotypes and improved representation of non-Western cultures. However, marginalised cultures remain underrepresented in mainstream models, and the absence of a unified standard for cultural competence leads to inconsistent cross-cultural performance.

4. Text to video generation

4.1. Research progress in technical aspects

4.1.1. Long video generation

Researchers solve key bottlenecks in long video generation - unstable temporal coherence, low production efficiency, and poor multimodal adaptability - through new frameworks that advance industrial use in film, gaming, and short-form content.

Picsart AI Research's StreamingT2V [20] introduces a streaming autoregressive diffusion framework supporting 80–1200+ frame (~40–50s) videos. Its CAM and APM modules, together with random mixing, ensure smooth transitions and infinite extension, outperforming peers in VBench temporal consistency.

Microsoft Research Asia's ARLON [21] combines AR and DiT models. It uses latent VQ-VAE compression and adaptive semantic injection, reducing denoising steps from 30 to 5–10 (6x faster) and outperforming OpenSora-V1.2 in 8/11 VBench metrics.

Zhejiang University and Oxford's FlexiFilm [22] is the first multimodal-conditioned long-video framework. Its temporal conditioner links text/image inputs with on-demand guidance, and resampling mitigates overexposure, producing high-quality 200+ frame (~8s) videos.

These innovations enable long video synthesis. However, ultra-long sequences may experience gradual quality degradation or repetitive content, and maintaining coherence at high resolution still requires substantial computational resources.

4.1.2. Controllable video generation from text

Researchers introduce new methods to solve key control problems in text-to-video generation, enabling image guidance, camera motion control, and human pose control for professional video making.

Mitsubishi Electric Research Laboratories's TI2V-Zero [23] uses a zero-shot, tuning-free image-conditioned method. It requires no extra training, integrates reference image features into video generation, and supports guidance insertion at any point for precise control and style transfer.

Tsinghua University and Stanford University's CameraCtrl [24] achieves precise camera pose control with parameterised trajectories and a plug-and-play module. It uses point trajectories instead of extrinsic matrices, enabling professional camera movements and giving users AI director-level control.

SenseTime and the Chinese University of Hong Kong's Follow-Your-Pose [25] realises pose-guided human video generation with a two-stage method. It builds a pose-controllable generator using text-pose pairs and optimises temporal attention with pose-free datasets, letting users control motion via stick figures or key points.

Zero-shot image guidance and parameterised motion control allow users to customise professional videos without extra training, giving them "director-like" scene control. However, complex camera movements or non-human pose control remain unreliable, and dynamic tracking may lose smoothness in some cases.

4.1.3. Physical perception generation

Traditional text-to-video models often lack physical realism and may violate basic physical laws. Recent studies address this through LLM-guided optimisation, fine-tuning of pre-trained models, and physics-parameter control, giving AI basic physical reasoning capabilities.

PhyT2V [26] improves physical realism through LLM-guided iterative self-refinement. It uses LLMs to analyse physical logic in prompts, generates physics-consistent guidance signals to correct unrealistic content, and increases physical realism by 2.3 times.

DiffPhy [27] fine-tunes pre-trained video diffusion models for physically accurate results. It extracts physical details (mass, collisions) from text via multimodal LLMs, integrates them into diffusion, and adheres to Newtonian laws without requiring additional training data.

PhysCtrl [28] is a physics-parameter and force-controlled image-to-video framework. It represents dynamics as 3D point trajectories, trains on 550K physics simulator datasets, and uses a spatiotemporal attention block to enforce constraints, enabling physics-compliant videos from a single image.

LLM-guided refinement and physics-parameter integration have significantly improved physical realism, ensuring compliance with basic physical laws for industrial and simulation use. However, complex interactions like fluid dynamics or soft-body collisions remain challenging, and these models require large-scale physics datasets, increasing training costs.

4.2. Application of text-to-video generation

4.2.1. Gaming

Traditional game content production involves high costs, long development cycles, and difficulty balancing interactivity with scene adaptability. Researchers improve workflows through custom designs that combine pre-trained models with game-specific data.

GameFactory [29] converts text inputs into interactive game videos. It achieves precise keyboard/mouse control via the GF-Minecraft dataset and a dedicated module, and realises scene generalisation by separating action control from style, enabling rapid prototyping and cost reduction.

Tencent's Hunyuan-GameCraft [30] is a high-dynamic interactive framework based on Hunyuan video technology. It unifies inputs for precise control, uses hybrid history conditioning for long-term consistency, and improves efficiency through model distillation. It runs in real time on an RTX 4090, reducing AAA game content production from weeks to hours.

GameGen-X [31] is the first diffusion Transformer for open-world game video generation. Through two-stage training, it learns game engine characteristics and integrates multimodal signals via InstructNet, allowing real-time player inputs to shape AAA-level open-world videos.

T2V technologies accelerate game prototyping and enable real-time interactive content generation, significantly shortening production cycles while supporting open-world adaptability. However, generated assets may lack the uniqueness of hand-crafted content, and real-time performance on low-end hardware remains a barrier to mainstream deployment.

4.2.2. Healthcare

Traditional medical videos are difficult to obtain, scarce, and risky for hands-on training. The following models generate realistic, controllable medical videos to aid training, education, and communication.

SurgSora [32] creates controllable surgical videos using anatomical images and motion cues. Its decoupled RGB-D/optical flow structure, Dual Semantic Injector, and Trajectory Controller address the shortage of medical videos, providing safe simulations for surgeon training and preoperative planning.

Endora [33] is the first endoscopic video generation model, combining a spatiotemporal Transformer with DINO vision priors. It establishes an endoscopy video benchmark, offering realistic videos for doctor training and surgical navigation to fix data scarcity and high-risk constraints.

WISA [34], uses a hierarchical physical knowledge decomposition framework for physics-aware text-to-video generation. It incorporates Mixture-of-Physical-Experts Attention (MoPA) and physical embedding injection - achieving 28.6% higher physical consistency scores, while maintaining only 3.5% parameter increase and 5% inference time overhead for industrial simulation applications.

Controllable surgical and endoscopic videos provide safe, accessible training tools, addressing the scarcity and risk associated with real medical training materials. However, the realism of dynamic surgical procedures still needs improvement, and integration with existing clinical workflow systems such as EHRs remains underdeveloped.

4.2.3. Industry and simulation

Industrial scenarios require high physical precision, fast inference, and practical deployment, while traditional solutions are costly and inefficient. Researchers address these challenges with physics-aware designs and multi-modal fusion.

Step-Video-Tl2V [35] generates high-quality industrial simulation videos by combining text with physics constraints such as friction coefficients. Its multi-scale physics-field decoupling structure achieves less than 2% inter-frame physical error, running 100× faster than traditional Computational Fluid Dynamics (CFD) in automotive cooling system simulation with 7% higher accuracy.

GR-2 [36] is a generative video-language-action model pre-trained on 38M videos and 50B tokens. Given an image and an instruction, it predicts industrial robot operation videos and generates control trajectories, achieving a 97.7% success rate across 100+ scenarios to boost smart factory efficiency.

CausVid [37] combines diffusion models' global modelling with autoregressive models' fast inference through a "teacher-student" structure. It generates industrial equipment videos 100× faster than traditional diffusion models, accelerating production line debugging.

Physics-constrained models provide high-precision simulation videos with fast inference, outperforming traditional methods in efficiency and accuracy. However, industrial scenarios often need domain-specific fine-tuning, and simulating rare faults in complex systems remains challenging.

5. Discussion

This review summarises the development of diffusion-based T2I and T2V technologies, with key findings aligning with recent studies. In T2I, innovations such as linear attention have overcome complexity constraints, demonstrating that latent space optimisation can balance resolution and efficiency - consistent with this review's "efficient module-application" framework [1]. In T2V, advances in physics-aware generation further confirm the importance of cross-modal reasoning in addressing physical realism challenges [26].

Key difficulties remain, including weak multilingual alignment, limited coherence in long-video generation, as well as prevalent semantic alignment gaps, spatio-temporal inconsistency collapse, and computational resource bottlenecks that plague both Text-to-Image (T2I) and Text-to-Video (T2V) generation [38]. Meanwhile, AI's growing ability to accelerate technical module integration—such as hierarchical representation learning and LLM-aided control strategies proposed in recent research [39]—opens new innovation pathways but requires deeper exploration to address the aforementioned limitations comprehensively.

Future research should prioritise three directions: leveraging image diffusion priors to unify styles in T2I-vector generation and reduce style drift [40]; constructing multicultural datasets to address the underrepresentation of non-Western content; and using LLMs to strengthen physical rule embedding for more rational dynamic scene generation. The framework proposed in this review can support these efforts, promoting fairer and more efficient industrialisation of generative models.

6. Conclusion

This paper reviews diffusion-based T2I and T2V technologies, which have become central to text-driven visual generation since DDPM surpassed GANs. Key breakthroughs such as Stable Diffusion's latent-space computation and DALL·E 2's CLIP-LLM integration have resolved major technical limitations. T2I has progressed in efficient high-resolution synthesis, multilingual rendering, and interactive control, while T2V has advanced in long-video generation, controllable synthesis, and physics-aware modelling. Both technologies now support applications in healthcare, gaming, animation, and cultural communication.

The proposed "Technical Module-Application-Evaluation" framework links technical innovations with real-world scenarios, highlighting understudied areas such as cultural competence. Despite substantial progress, challenges persist, including quality-efficiency trade-offs, underrepresentation of non-Western cultures, and insufficient physical realism. Future research should focus on balancing quality and efficiency, building multicultural datasets, and deepening LLM-diffusion integration to support fair and reliable industrialisation.

References

[1] Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., & Han, S. (2025). SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. *International Conference on Learning Representations*, 1-25.

[2] Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., & Li, Z. (2024). *PixArt-Σ: Weak-to-strong training of diffusion transformer for 4K text-to-image generation*. arXiv. https: //arxiv.org/abs/2403.04692v2

[3] Yu, Y., Zhang, H., Zhang, Z., Lin, Z., Zhang, J., & Ding, C. (2025). *TurboFill: Adapting few-step text-to-image models for fast image inpainting*. arXiv. https: //arxiv.org/abs/2504.00996

[4] Beyer, L., Li, T., Chen, X., Karaman, S., & He, K. (2025). *Highly compressed tokenizer can generate without training*. arXiv. https: //arxiv.org/abs/2506.08257

[5] ByteDance Seed Team. (2025). *Seedream 3.0: A high-performance Chinese-English bilingual image generation foundation model* (Technical Report). ByteDance.

[6] Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., … (2025). *Qwen-Imag technical report*. arXiv. https: //arxiv.org/abs/2508.02324

[7] Wang, Y., Han, C., Li, Y., Jin, Z., Li, X., Du, S., Tao, W., Li, S., Yang, Y., Yuan, C., & Lin, L. (2025). UniGlyph: Unified segmentation-conditioned diffusion for precise visual text synthesis. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 18335–18344.

[8] Qian, G., Zhang, R., Chen, T., Dalva, Y., Goyal, A., Menapace, W., Skorokhodov, I., Dong, M., Sahni, A., Ostashev, D., Hu, J., Tulyakov, S., & Wang, K. (2025). *LayerComposer: Interactive personalized T2I via spatially-aware layered canvas*. arXiv. https: //arxiv.org/abs/2510.20820

[9] Hu, X., Wang, J., Chen, H., Zhang, W., Wang, B., Li, Y., & Nan, H. (2025). DynamicID: Zero-shot multi-ID image personalization with flexible facial editability. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 789–798.

[10] Chen, C., Shi, M., Zhang, G., & Shi, H. (2025). T2I-Copilot: A training-free multi-agent text-to-image system for enhanced prompt interpretation and interactive generation. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 19396–19405.

[11] Zhang, Z., Yu, Y., Chen, Y., Yang, X., & Yeo, S. (2025). MedUnifier: Unifying vision-and-language pre-training on medical data with vision generation task using discrete visual representations. *2025 IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), 10215–10224.

[12] Bhat, A., Bose, R., Nwoye, C., & Padoy, N. (2025). SimGen: A diffusion-based framework for simultaneous surgical image and segmentation mask generation. *Medical Image Analysis, 88*, 103987.

[13] Dai, F., Yao, S., Wang, M., Zhu, Y., Qiu, X., Sun, P., Qiu, C., Yin, J., Shen, G., Sun, J., Wang, M., Wang, Y., Yang, Z., Sang, J., Wang, X., Sun, F., Cai, W., Zhang, X., & Lu, H. (2025). Improving AI models for rare thyroid cancer subtype by text guided diffusion models. *Nature Communications, 16*, 4449. https: //doi.org/10.1038/s41467-025-59478-8

[14] Zheng, H., Zhang, W., Wang, Y., Li, J., Lv, Z., Min, X., Li, M., Zhang, D., Tang, S., & Zhuang, Y. (2024). *LASER: Tuning-free LLM-driven attention control for efficient text-conditioned image-to-animation*. arXiv. https: //arxiv.org/abs/2404.13558v3

[15] Jian, J., Tian, C., Wang, J., Zhang, Y., Li, X., & Tong, Y. (2025). *DiffSensei: Bridging multi-modal LLMs and diffusion models for customized manga generation*. arXiv. https: //arxiv.org/abs/2412.07589v2

[16] Yang, Y., Fan, L., Lin, Z., Wang, F., & Zhang, Z. (2025). *LayerAnimate: Layer-level control for animation*. arXiv. https: //arxiv.org/abs/2501.08295v3

[17] Bayramli, Z., Suleymanzade, A., An, N., Ahmad, H., Kim, E., Park, J., Thorne, J., & Oh, A. (2025). Diffusion models through a global lens: Are they culturally inclusive? *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 31137–31155.

[18] Kannen, N., Ahmad, A., Andreetto, M., Prabhakaran, V., Prabhu, U., Dieng, A., Bhattacharyya, P., & Dave, S. (2025). *Beyond aesthetics: Cultural competence in text-to-image models*. arXiv. https: //arxiv.org/abs/2407.06863v6

[19] Nayak, S., Bhatia, M., Zhang, X., Rieser, V., Hendricks, L., van Steenkiste, S., Goyal, Y., Stańczak, K., & Agrawal, A. (2025). *CulturalFrames: Assessing cultural expectation alignment in text-to-image models and evaluation metrics*. arXiv. https: //arxiv.org/abs/2506.08835v2

[20] Henschel, R., Khachatryan, L., Poghosyan, H., Hayrapetyan, D., Tadevosyan, V., Wang, Z., Navasardyan, S., & Shi, H. (2025). StreamingT2V: Consistent, dynamic, and extendable long video generation from text. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2568–2577.

[21] Li, Z., Hu, S., Liu, S., Zhou, L., Choi, J., Meng, L., Guo, X., Li, J., Ling, H., & Wei, F. (2025). ARLON: Boosting diffusion transformers with autoregressive models for long video generation. *International Conference on Learning Representations*, 912–923.

[22] Ouyang, Y., Yuan, J., Zhao, H., Wang, G., & Zhao, B. (2024). *FlexiFilm: Long video generation with flexible conditions*. arXiv. https: //arxiv.org/abs/2404.18620

[23] Ni, H., Egger, B., Lohit, S., Cherian, A., Wang, Y., Koike-Akino, T., Huang, S., & Marks, T. (2024). TI2V-Zero: Zero-shot image conditioning for text-to-video diffusion models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9015–9025.

[24] He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., & Yang, C. (2024). *CameraCtrl: Enabling camera control for text-to-video generation*. arXiv. https: //arxiv.org/abs/2404.02101

[25] Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., & Chen, Q. (2024). Follow-Your-Pose: Pose-guided text-to-video generation using pose-free videos. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(5), 4117–4125. https: //doi.org/10.1609/aaai.v38i5.28206

[26] Xue, Q., Yin, X., Yang, B., & Gao, W. (2025). PhyT2V: LLM-guided iterative self-refinement for physics-grounded text-to-video generation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18826–18836.

[27] Zhang, K., Xiao, C., Xu, J., Mei, Y., & Patel, V. (2025). *Think before you diffuse: LLMs-guided physics-aware video generation.* arXiv. https: //arxiv.org/abs/2505.21653v3

[28] Wang, C., Chen, C., Huang, Y., Dou, Z., Liu, Y., Gu, J., & Liu, L. (2024). *PhysCtrl: Generative physics for controllable and physics-grounded video generation*. arXiv. https: //arxiv.org/abs/2509.20358

[29] Yu, J., Qin, Y., Wang, X., Wan, P., Zhang, D., & Liu, X. (2025). *GameFactory: Creating new games with generative interactive videos.* arXiv. https: //arxiv.org/abs/2501.08325v4

[30] Li, J., Tang, J., Xu, Z., Wu, L., Zhou, Y., Shao, S., Yu, T., Cao, Z., & Lu, Q. (2025). *Hunyuan-GameCraft: High-dynamic interactive game video generation with hybrid history condition*. arXiv. https: //arxiv.org/abs/2506.17201

[31] Che, H., He, X., Liu, Q., Jin, C., & Chen, H. (2024). *GameGen-X: Interactive open-world game video generation.* arXiv. https: //arxiv.org/abs/2411.00769v3

[32] Chen, T., Yang, S., Wang, J., Bai, L., Ren, H., & Zhou, L. (2024). *SurgSora: Decoupled RGBD-flow diffusion model for controllable surgical video generation*. arXiv. https: //arxiv.org/abs/2412.14018v3

[33] Li, C., Liu, H., Liu, Y., Feng, B., Li, W., Liu, X., Chen, Z., Shao, J., & Yuan, Y. (2024). Endora: Video generation models as endoscopy simulators. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 230–240.

[34] Wang, R., Chen, J., Ji, K., Cai, Z., Chen, S., Yang, Y., & Wang, B. (2025). *MedGen: Unlocking medical video generation by scaling granularly-annotated medical videos*. arXiv. https: //arxiv.org/abs/2507.05675

[35] StepFun Team. (2025). *Step-Video-TI2V technical report: A state-of-the-art text-driven image-to-video generation model.* arXiv. https: //arxiv.org/abs/2503.11251

[36] Cheang, C., Chen, G., Jing, Y., Kong, T., Li, H., Li, Y., Liu, Y., Wu, H., Xu, J., Yang, Y., Zhang, H., & Zhu, M. (2024). *GR-2: A generative video-language-action model with web-scale knowledge for industrial robotics*. arXiv. https: //arxiv.org/abs/2410.06158

[37] Wang, J., Ma, A., Cao, K., Zheng, J., Zhang, Z., Feng, J., Liu, S., Ma, Y., Cheng, B., Leng, D., Yin, Y., & Liang, X. (2025). *WISA: World simulator assistant for physics-aware text-to-video generation*. arXiv. https: //arxiv.org/abs/2503.08153v1

[38] Kumar, N., Bhandari, P., & Maragatham, G. (2025). *Bridging text and video generation: A survey*. arXiv. https: //arxiv.org/abs/2510.04999v1

[39] Zhao, C., Liu, M., Wang, W., Chen, W., Wang, F., Chen, H., Zhang, B., & Shen, C. (2025). *MovieDreamer: Hierarchical generation for coherent long visual sequence*. arXiv. https: //arxiv.org/abs/2407.16655v3

[40] Zhang, P., Zhao, N., & Liao, J. (2025). Style customization of text-to-vector generation with image diffusion priors. *ACM SIGGRAPH 2025 Conference Proceedings, 72*, 1–11.