Review

Generative AI for Text-to-Video Generation: Recent Advances and Future Directions

Kadhim Hayawi and Sakib Shahriar *

Posted Date: 17 December 2025

doi: 10.20944/preprints202512.1476.v1

Keywords: large language model; video generation; text-to-video generation; literature review

 doi:10.20944/preprints202512.1476.v1

Review

Generative AI for Text-to-Video Generation: Recent Advances and Future Directions

Kadhim Hayawi [1] and Sakib Shahriar [2,*]

[1] College of Interdisciplinary Studies, Zayed University, Dubai, United Arab Emirates
[2] School of Computer Science, University of Guelph, Guelph, ON, Canada
* Correspondence: shahrias@uoguelph.ca

Abstract

Text-to-video (T2V) generation has recently emerged as a transformative technology within the field of generative AI, enabling the creation of realistic, temporally coherent videos based on natural language descriptions. This paradigm provides significant added value in many domains such as creative media, human-computer interaction, immersive learning, and simulation. Despite its growing importance, systematic discussion of T2V is still limited compared with adjacent modalities such as text-to-image and image-to-video. To alleviate the scarcity of discussions in the T2V field, this paper provides a systematic review of works published from 2024 onward, consolidating fragmented contributions across the field. We survey and categorize the selected literature into three principal areas, namely, T2V methods, datasets, and evaluation practices, and further subdivide each area into subcategories that reflect recurring themes and methodological patterns in the literature. Emphasis will then be placed on identifying key research opportunities and open challenges that need further investigation.

Keywords: large language model; video generation; text-to-video generation; literature review

1. Introduction

Recent advances in generative artificial intelligence (AI) are reshaping the production of multimedia content, including text, images, music, and video. A promising frontier is **text-to-video (T2V) generation**, which has gained growing popularity in recent years. Video synthesis has become the core of this popularity since it enables the creation of realistic videos with high visual fidelity and coherent temporal transitions. Among the technological frameworks that supported this progress are diffusion models that have shown remarkable success in generating high-quality video content, exemplified by Sora's groundbreaking video generation [1], which demonstrates the performance of Diffusion Transformer (DiT) architectures. Moreover, the emergence of alternative generative paradigms like autoregressive models (e.g., Video Pixel Networks [2]) and variational autoencoders (VAEs) (e.g., Stochastic Video Generation (SVG) [3]) has also advanced this field. Whether trained from scratch or fine-tuned from pre-existing architectures, these models have rejuvenated research in T2V generation.

Alongside synthesis, video understanding is essential for endowing models with interpretation, evaluation, and the ability to enhance generated videos. An effective video understanding ensures semantic alignment between the generated video and the input text, thereby maintaining narrative coherence and temporal logic. Integrating Large Language Models (LLMs) like GPT [4] or LLaMA [5,6] with visual encoders [7,8] in vision-and-language modeling have significantly improved performance on video understanding tasks such as caption generation, action recognition, and temporal event localization. Another pillar in this field is the objective assessment and benchmarking to evaluate key aspects of the generated videos like visual realism, text-video alignment, and alignment with human preferences, which enables to measure and guide progress in T2V research. Another active research direction involves developing benchmark datasets, ranging from short-form video collections to more complex and prompt-driven datasets.

 doi:10.20944/preprints202512.1476.v1

In order to enrich the literature on text-to-video generation, this paper provides a comprehensive analysis of the potential of Generative AI and LLMs in video generation along with an overview of video evaluation frameworks and benchmark datasets. Focusing on studies published from 2024 onward, we review the current state of the field through the lens of the following questions:

- RQ1: What are the key technological advances that have driven progress in the aforementioned video research fields?
- RQ2: What are the current challenges and best practices in evaluating T2V models, and how do benchmark datasets support the development of robust text-to-video generators?
- RQ3: What are the primary technical challenges and future research directions in leveraging Generative AI and LLMs for text-to-video generation?

The paper is structured as follows. Section 2 presents an overview of LLMs and generative AI, including variational autoencoders (VAEs), autoregressive models, and Diffusion Transformers (DiTs), as well as the pre-training and transfer learning paradigms that underpin them. In Section 3, selected relevant papers, statistical analyses of the selected literature are presented, covering publication types, industrial sectors, shop typology, and other relevant factors. Section 4 presents a classification and literature review of the selected publications. Section 6 discusses the main findings derived from the reviewed papers and identifies many future research needs.

2. Background

It is imperative to understand the background of generative AI and LLMs to contextualize their performance and versatility in text-to-video generation. This section presents a brief discussion of generative AI and LLM architectures and their main characteristics. Moreover, it briefly covers the fundamental learning paradigms that support these models, namely pretraining and transfer learning.

2.1. Large Language Models and Generative Architectures

Generative AI has enabled machines to generate coherent, high-quality outputs across many modalities, such as text, images, audio, and video. Within this framework, deep learning-based generative models, including diffusion-based models, autoregressive models, and variational autoencoders (VAEs), have shown high performance in generating realistic, multimodal content across a wide range of applications. In order to generate new, realistic samples that mimic those from the original dataset, generative AI models attempt to learn the underlying data distribution of a given modality.

Variational Autoencoders (VAEs) are generative models that use two networks: an encoder network and a decoder network. While the encoder network learns to approximate the posterior distribution of latent variables, the decoder network learns to reconstruct the original input from the latent representation. This framework enables VAEs to generate samples that are consistent with the underlying data distribution by optimizing a variational lower bound on the data likelihood. In the context of video generation, VAEs have been extended to work with spatiotemporal data by modeling temporal dependencies within the latent space by using temporal transformers or recurrent layers for sequence modeling, or by introducing hierarchical latent structures. For example, the Stochastic Video Generation (SVG) framework [3] uses recurrent latent variables to capture the inherent uncertainty and dynamics in video sequences, thereby enabling the generation of multiple futures conditioned on previous frames.

Diffusion models are another type of generative models that have recently attracted increasing attention in video generation. Unlike VAEs and autoregressive models, diffusion models generate data through a gradual denoising process. These models learn to reverse a fixed noising procedure that incrementally corrupts the data—typically by adding Gaussian noise over multiple steps—until it becomes pure noise. This denoising process is often parameterized by conditional neural networks—commonly U-Nets or transformer architectures—that predict either the noise added at each step or the clean data itself. Training is generally performed by optimizing a denoising score-matching objective. In video generation, diffusion models have been extended to capture both spatial and temporal correlations

 doi:10.20944/preprints202512.1476.v1

through various architectural designs. For instance, Video Diffusion Models (VDMs) [9] incorporate 3D convolutional denoising networks, as well as motion decoders and temporal attention modules, enabling the model to operate on video clips as spatiotemporal volumes.

Autoregressive models are another salient class of generative models that create data one element at a time by modeling the conditional probability of each element given the preceding ones. These models are trained to maximize the likelihood of observed sequences, typically by minimizing their negative log-likelihood. In video generation, autoregressive models usually synthesize video frames in a fixed temporal order, frame by frame or pixel by pixel, by using previously generated content as a guide. Modern autoregressive video models frequently incorporate temporal transformers and attention layers to effectively model long-range temporal relationships and dependencies among frames. A notable example includes Video Pixel Networks (VPNs) [2], which decomposes video generation into conditional distributions over spatiotemporal patches. Architecturally, VPNs extend PixelCNN [10] with convolutional Long Short-Term Memory (LSTM) to model both spatial and temporal dependencies. Specifically, VPNs factorize video distributions into a four-dimensional dependency chain—corresponding to spatial width, spatial height, temporal frames, and color channels. Each pixel in a frame is generated based on previously generated pixels within the same frame, as well as pixels from earlier frames, thereby capturing both intra-frame and inter-frame correlations.

2.2. Pre-training and Transfer Learning Paradigms

In order to generate temporally coherent video sequences, text-to-video (T2V) models are commonly trained using supervised learning on large-scale captioned video datasets, such as LAION-5B [11], Panda-70M [12], and WebVid-10M [13]. In these corpora, video-text examples serve as labeled data guiding the model to learn how to map textual descriptions (prompts) to corresponding visual and temporal features. Two main training paradigms are widely used:

- **Training from Scratch:** The model weights are randomly initialized, and all spatial and temporal patterns are learned entirely from large-scale paired video–text datasets. This method enables full flexibility when designing model architectures (e.g., spatiotemporal transformers or 3D convolutional networks) [14].
- **Fine-Tuning:** In this paradigm, the model is initialized with pre-trained weights and then adapted to the video domain through further training on paired video–text datasets. Fine-tuning involves updating the entire model or selectively updating specific modules (e.g., temporal transformers or attention layers). Techniques such as lightweight adapters, Low-Rank Adaptation (LoRA) layers [15], or partial-freeze schedules [16] (e.g., freezing the text encoder and early spatial layers while updating temporal transformer blocks) are often used to reduce computational cost.

Due to the high computational cost and data requirements of training from scratch and the growing availability of pre-trained T2V backbones, fine-tuning has emerged as the most popular method. Furthermore, the spatial and semantic priors encoded in the pretrained models enable faster convergence, more efficient adaptation to temporal dynamics, and high-quality video generation with fewer resources [17].

3. Bibliometric Analysis

As stated in Section 1, the objective of the present research is to scrutinize the current status of research on T2V generation and identify potential research directions. Research questions were formulated in Section 1. Research question 1 (**RQ1**) aims to identify the main innovations that have accelerated progress in T2V, including advances in T2V models and video–text datasets. Research question 2 (**RQ2**) examines the evaluation landscape for T2V models, focusing on both persistent challenges and emerging best practices. Finally, research question 3 (**RQ3**) aims to identify the challenges and future research opportunities in T2V generation.

In order to gather a comprehensive range of publications, data were collected from prominent academic sources, including Google Scholar and arXiv. All English-language publications identified as

 doi:10.20944/preprints202512.1476.v1

journal articles, conference proceedings, or preprints were included, without imposing restrictions on publication venue or journal ranking. After removing duplicates and excluding papers that were not directly relevant to the research scope, a total of 69 publications were retained, which are listed in Table 1.

To deepen our understanding of the evolving landscape of T2V research, we categorized and analyzed the 69 selected publications along several dimensions. These statistics provide insight into key trends, common practices, and gaps in the field, as illustrated in Figures 4–6.

Regarding publication types (Figure 1), conference papers dominate (72%), with preprints/tech reports at 21%, while peer-reviewed journal articles present only 7%. Figure 2 categorizes papers by research theme. More than three-quarters (77%) of the corpus focuses on *video synthesis* directly. In contrast, only 18% focus on *objective assessment & benchmarks*, and a mere 5% on *video understanding*.

Figure 3 shows that an encouraging 91% of the surveyed papers publicly release their code, reflecting the open research culture that has similarly accelerated progress in T2V generation models. However, this openness contrasts sharply with the limited accessibility of large-scale video datasets. The remaining 9% are typically corporate tech reports with either (i) proprietary data pipelines or (ii) safety concerns around video content. Figure 4 summarizes the training strategies adopted in the surveyed models. Almost two-thirds of recent T2V work (68%) relies on *fine-tuning existing T2V models*. Meanwhile, 21% of works still pursue training from scratch, especially when introducing novel T2V architectures. Specifically, 11% adopt *training-free* strategies, indicating a growing interest in lightweight, low-compute alternatives appropriate for rapid experimentation and deployment.

In terms of architectural design (Figure 5), we observe the dominance of diffusion models, often enhanced by transformer backbones. This reflects the field's convergence on scalable, generative architectures capable of producing high-fidelity, temporally consistent outputs. Other architectures include GAN-based models and hybrid transformer–CNN models, yet these are becoming less common.

Finally, Figure 6 shows the distribution of dataset types that have been used in the selected literature. More than half of the studies (54%) leveraged *large-scale, web-scraped video–text corpora* such as WebVid-10M or Panda-70M. Domain-specific collections (15%) and benchmark/evaluation suites (14%) together account for under one-third, while *synthetic/LLM-generated* video–text pairs appear in 10%. High-quality, fully curated datasets occupy only 7%.

Table 1. Selected literature.

	Authors	Publication title
1	Bahmani et al. [18]	AC3D: Analyzing and Improving 3D Camera Control in Video Diffusion Transformers.
2	Bao et al. [19]	Vidu: a Highly Consistent, Dynamic and Skilled Text-to-Video Generator with Diffusion Models.
3	Cai et al. [20]	DiTCtrl: Exploring Attention Control in Multi-Modal Diffusion Transformer for Tuning-Free Multi-Prompt Longer Video Generation.
4	Chen et al. [21]	ShareGPT4Video: Improving Video Understanding and Generation with Better Captions.
5	Chen et al. [22]	VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models.
6	Choi et al. [23]	We'll Fix it in Post: Improving Text-to-Video Generation with Neuro-Symbolic Feedback.
7	Cuttano et al. [24]	SAMWISE: Infusing Wisdom in SAM2 for Text-Driven Video Segmentation.
8	Dalal et al. [25]	One-Minute Video Generation with Test-Time Training.

Continued on next page

 doi:10.20944/preprints202512.1476.v1

	Authors	Publication title
9	Fei et al. [26]	Dysen-VDM: Empowering Dynamics-aware Text-to-Video Diffusion with LLMs.
10	Gal et al. [27]	Breathing Life Into Sketches Using Text-to-Video Priors.
11	Girdhar et al. [28]	Factorizing Text-to-Video Generation by Explicit Image Conditioning.
12	Guo et al. [29]	SparseCtrl: Adding Sparse Controls to Text-to-Video Diffusion Models.
13	Guo et al. [30]	Can You Count to Nine? A Human Evaluation Benchmark for Counting Limits in Modern Text-to-Video Models.
14	Guo et al. [31]	T2VPhysBench: A First-Principles Benchmark for Physical Consistency in Text-to-Video Generation.
15	Henschel et al. [32]	StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text.
16	Huang et al. [33]	VBench: Comprehensive Benchmark Suite for Video Generative Models.
17	Jeong et al. [34]	VMC: Video Motion Customization using Temporal Attention Adaption for Text-to-Video Diffusion Models.
18	Jiang et al. [35]	VideoBooth: Diffusion-based Video Generation with Image Prompts.
19	Ju et al. [36]	MiraData: A Large-Scale Video Dataset with Long Durations and Structured Captions.
20	Kou et al. [37]	Subjective-Aligned Dataset and Metric for Text-to-Video Quality Assessment.
21	Li et al. [38]	PhyT2V: LLM-Guided Iterative Self-Refinement for Physics-Grounded Text-to-Video Generation
22	Li et al. [39]	Training-free Guidance in Text-to-Video Generation via Multimodal Planning and Structured Noise Initialization.
23	Liao et al. [40]	Evaluation of Text-to-Video Generation Models: A Dynamics Perspective.
24	Lin et al. [41]	Open-Sora Plan: Open-Source Large Video Generation Model.
25	Liu et al. [42]	VideoDPO: Omni-Preference Alignment for Video Diffusion Generation.
26	Liu et al. [43]	Timestep Embedding Tells: It's Time to Cache for Video Diffusion Model.
27	Lv et al. [44]	GPT4Motion: Scripting Physical Motions in Text-to-Video Generation via Blender-Oriented GPT Planning.
28	Ma et al. [45]	Follow Your Pose: Pose-Guided Text-to-Video Generation Using Pose-Free Videos.
29	Menapace et al. [46]	Snap Video: Scaled Spatiotemporal Transformers for Text-to-Video Synthesis.
30	Miao et al. [47]	T2VSafetyBench: Evaluating the Safety of Text-to-Video Generative Models.
31	Mohamed and Lucke-Wold [48]	Text-to-Video Generative Artificial Intelligence: Sora in Neurosurgery.
32	Nan et al. [49]	OpenVid-1M: A Large-Scale High-Quality Dataset for Text-to-Video Generation.

Continued on next page

 doi:10.20944/preprints202512.1476.v1

	Authors	Publication title
33	Qin et al. [50]	xGen-VideoSyn-1: High-Fidelity Text-to-Video Synthesis with Compressed Representations.
34	Qing et al. [51]	Hierarchical Spatio-temporal Decoupling for Text-to-Video Generation.
35	Qu et al. [52]	Exploring AIGC Video Quality: A Focus on Visual Harmony Video-Text Consistency and Domain Distribution Gap.
36	Rawte et al. [53]	ViBe: A Text-to-Video Benchmark for Evaluating Hallucination in Large Multimodal Models.
37	Ren et al. [54]	Customize-A-Video: One-Shot Motion Customization of Text-to-Video Diffusion Models.
38	Sharan et al. [55]	Neuro-Symbolic Evaluation of Text-to-Video Models using Formal Verification.
39	Si et al. [56]	FreeU: Free Lunch in Diffusion U-Net.
40	Tan et al. [57]	Mimir: Improving Video Diffusion Models for Precise Text Understanding.
41	Tian et al. [58]	VideoTetris: Towards Compositional Text-to-Video Generation.
42	Wang et al. [59]	A Recipe for Scaling up Text-to-Video Generation with Text-free Videos.
43	Wang et al. [60]	LaVie: High-Quality Video Generation with Cascaded Latent Diffusion Models.
44	Wang and Yang [61]	VidProM: A Million-scale Real Prompt-Gallery Dataset for Text-to-Video Diffusion Models.
45	Sun et al. [62]	T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation.
46	Wang et al. [63]	LOVE: Benchmarking and Evaluating Text-to-Video Generation and Video-to-Text Interpretation.
47	Wang et al. [64]	T2VBench: Benchmarking Temporal Dynamics for Text-to-Video Generation
48	Wang et al. [65]	WISA: World Simulator Assistant for Physics-Aware Text-to-Video Generation.
49	Wei et al. [66]	DreamVideo: Composing Your Dream Videos with Customized Subject and Motion.
50	Weng et al. [67]	ART-V: Auto-Regressive Text-to-Video Generation with Diffusion Models.
51	Wu et al. [68]	Towards A Better Metric for Text-to-Video Generation.
52	Wu et al. [69]	DragAnything: Motion Control for Anything Using Entity Representation.
53	Xie et al. [70]	Progressive Autoregressive Video Diffusion Models.
54	Xing et al. [71]	DynamiCrafter: Animating Open-Domain Images with Video Diffusion Priors.
55	Yang et al. [72]	CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer.
56	Yang et al. [73]	IPO: Iterative Preference Optimization for Text-to-Video Generation.

Continued on next page

 doi:10.20944/preprints202512.1476.v1

	Authors	Publication title
57	Yin et al. [74]	From Slow Bidirectional to Fast Autoregressive Video Diffusion Models.
58	Yin et al. [75]	Step-Video-T2V Technical Report: The Practice, Challenges, and Future of Video Foundation Model.
59	Yuan et al. [76]	ChronoMagic-Bench: A Benchmark for Metamorphic Evaluation of Text-to-Time-lapse Video Generation.
60	Yuan et al. [77]	MagicTime: Time-lapse Video Generation Models as Metamoraphic Simulators.
61	Yuan et al. [78]	Inflation With Diffusion: Efficient Temporal Adaptation for Text-to-Video Super-Resolution.
62	Yuan et al. [79]	Identity-Preserving Text-to-Video Generation by Frequency Decomposition.
63	Zhang et al. [80]	CAMEL: CAusal Motion Enhancement Tailored for Lifting Text-driven Video Editing.
64	Zhang et al. [81]	Style-A-Video: Agile Diffusion for Arbitrary Text-Based Video Style Transfer.
65	Zhang et al. [82]	Tora: Trajectory-oriented Diffusion Transformer for Video Generation.
66	Zhang et al. [83]	Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation.
67	Zhao et al. [84]	MotionDirector: Motion Customization of Text-to-Video Diffusion Models.
68	Zhou et al. [85]	HiTVideo: Hierarchical Tokenizers for Enhancing Text-to-Video Generation with Autoregressive Large Language Models.
69	Zhu et al. [86]	Exploring Pre-trained Text-to-Video Diffusion Models for Referring Video Object Segmentation.

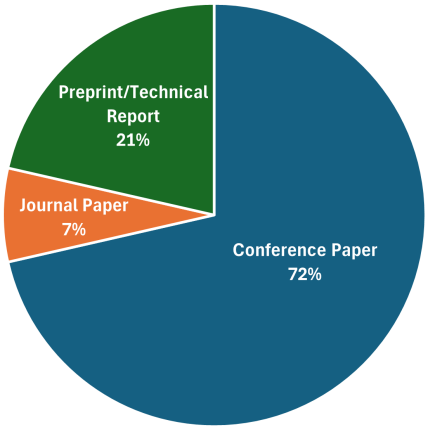

Figure 1. Types of publications.

 doi:10.20944/preprints202512.1476.v1

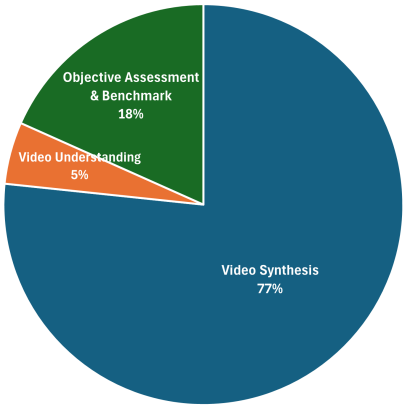

Figure 2. Research themes of publications.

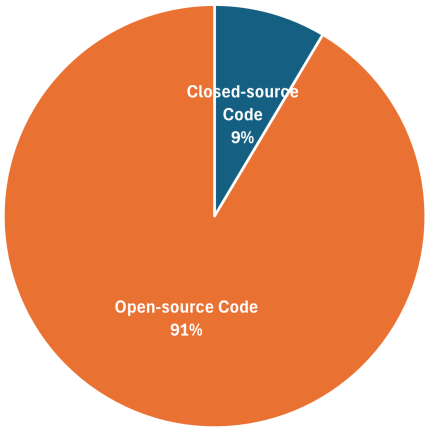

Figure 3. Open-source availability.

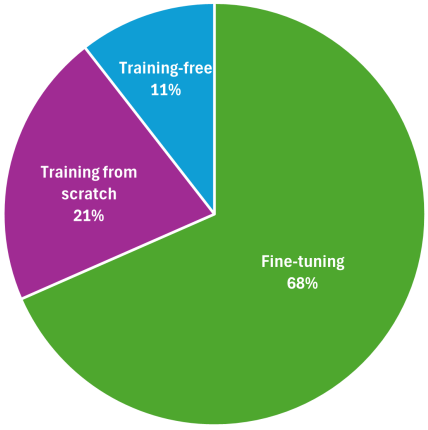

Figure 4. Training strategies.

 doi:10.20944/preprints202512.1476.v1

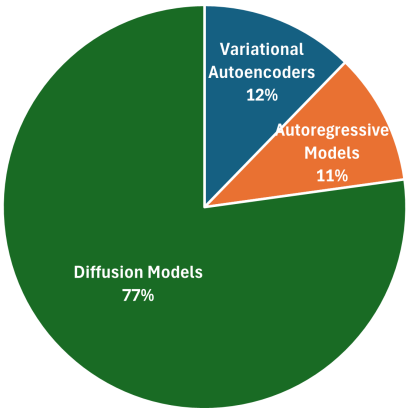

Figure 5. Architectural classification.

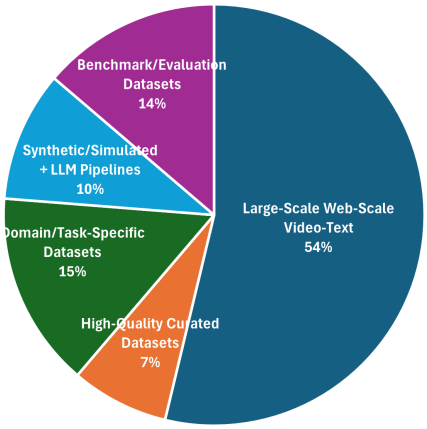

Figure 6. Types of datasets.

4. Literature Review

As outlined in Section 1, the scope of text-to-video research covers three primary themes: *video synthesis*, *video understanding*, and *objective assessment and benchmarking*. This section surveys and categorizes key contributions from the literature, structured around the following research questions:

RQ1: What are the key technological advances that have driven progress in the aforementioned video research fields?

RQ2: What are the current challenges and best practices in evaluating T2V models, and how do benchmark datasets support the development of robust text-to-video generators?

4.1. Video Synthesis

Driven by recent advances in transformer architectures (Vaswani, 2017; Peebles and Xie, 2023) and diffusion models (Ho et al., 2020; Song et al., 2020), visual content generation has shown remarkable capabilities when conditioned on text prompts. These developments have sparked widespread interest and a surge of new research efforts.

A notable milestone is the introduction of **Sora** [1], OpenAI's diffusion-transformer-based video generation model, which demonstrated striking results and marked a leap forward—ushering in a new era of text-driven video synthesis. A key innovation in Sora is its integration of a descriptive captioning model with GPT-based prompt expansion (adapted from DALL·E 3 [87]). This "recaptioning" strategy enriches training inputs with detailed textual annotations, improving alignment between user prompts and generated videos. Similar to other diffusion models, generation in Sora starts with a noisy grid of latent patches, which are iteratively denoised. This process is guided by the transformer model and conditioned on user-provided text to generate coherent and text-aligned videos.

Building on the promise of Sora, Mohamed and Lucke-Wold [48] investigates its applicability within the medical domain, particularly in neurosurgery. Their study emphasizes the promise of such

 doi:10.20944/preprints202512.1476.v1

models for medical education, surgical simulation, and patient communication by rapidly generating tailored procedural or anatomical videos from natural language prompts. However, they also identify several key challenges hindering clinical adoption, including physically implausible or anatomically inaccurate motion, generation of non-existent or irrelevant entities, unnatural object morphing, abrupt transitions, and flawed modeling of physical interactions.

Alongside Sora, other innovations have emerged, covering primarily four subcategories, namely, automatic T2V generation, controllable T2V generation, video style transfer & editing, and video quality & inference enhancement (see Figure 7).

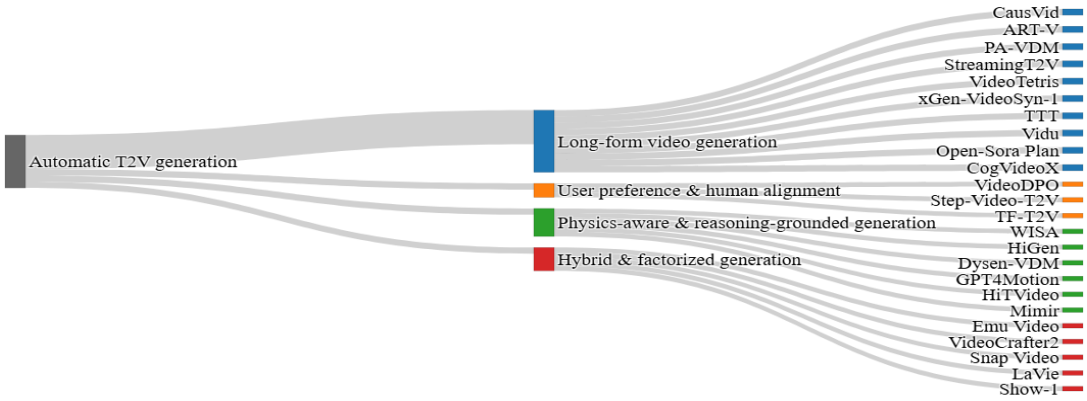

Figure 7. Schematic overview of T2V model classified within the automatic T2V generation category.

4.1.1. Automatic T2V Generation

In this category, related works can be further grouped into four sub-categories: long-form video generation, user preference & human alignment, physics-aware & reasoning-grounded generation, and hybrid & factorized generation.

a) Long-form video generation

Generating long videos poses unique challenges, including maintaining temporal coherence, semantic consistency, and computational efficiency across extended durations. Many works have addressed long-form T2V generation to move beyond a few-second clip toward narratives that unfold over tens of seconds or even minutes.

For instance, while many existing diffusion models rely on bidirectional attention mechanisms that require access to the entire sequence—including future frames—for generating each frame, Yin et al. [74] introduced **CausVid**, a video diffusion framework that re-engineers diffusion transformers into autoregressive transformers with causal attention. This enables frame-by-frame generation based only on preceding frames, enabling interactive and progressive synthesis. To accelerate inference, CausVid extends Distribution Matching Distillation (DMD) to video, reducing sampling steps from around 50 to just 4 via a teacher–student framework, where the original bidirectional model supervises the autoregressive student. Training is stabilized using an initialization scheme based on the teacher's continuous ODE trajectories, while an asymmetric distillation strategy mitigates error accumulation during generation. Additionally, key-value caching reuses attention states across frames, reducing redundant computation and significantly boosting speed. Despite being trained primarily on short clips, CausVid can synthesize long videos (~30 seconds at 640×352 resolution) with strong temporal consistency and achieves real-time performance (~9.4 FPS on a single GPU), making it suitable for low-latency applications.

Building on the autoregressive paradigm, Weng et al. [67] introduced **ART-V**, an auto-regressive (AR) T2V framework built on pretrained text-to-image diffusion models, extended for long video synthesis at resolutions up to 768×768. ART-V generates each frame sequentially, each conditioned on one or more previous frames, using a lightweight T2I-Adapter to incorporate temporal context. In order to mitigate

 doi:10.20944/preprints202512.1476.v1

appearance drifting—a common issue in AR video generation where content gradually deviates—ART-V employs a masked diffusion mechanism that learns to copy stable regions from reference frames and selectively generate new content where changes are needed, minimizing error accumulation. An anchored conditioning strategy further enhances global consistency in scene layout and appearance throughout the video by conditioning on a fixed anchor frame (user-defined or generated) alongside previous frames. To simulate real-world inference conditions and boost robustness, ART-V applies noise augmentation to reference frames during training. The framework supports diverse conditioning modes—text prompts, static images, or hybrids—enabling applications such as long-form video synthesis from sequential prompts and animation of still images with textual descriptions.

Further extending autoregressive techniques, Xie et al. [70] proposed **Progressive Autoregressive Video Diffusion Models (PA-VDM)**, an approach that extends existing video diffusion models to autoregressively generate high-quality, temporally coherent long videos—up to 60 seconds (1440 frames, at 24 FPS) at resolutions ranging from 176×320 to 240×424). PA-VDM introduces progressive noise scheduling, assigning noise levels per frame rather than uniformly across the entire video. Frames nearer to the current generation step are assigned lower noise (indicating higher certainty), while more distant frames receive higher noise, enabling smooth information propagation and temporal continuity. To further enhance coherence, the model adopts autoregressive frame generation with overlapping attention, progressively denoising small frame intervals instead of the entire sequence at once. This overlapping structure allows stronger correlations between adjacent frames and mitigates artifacts or abrupt transitions often seen in previous autoregressive approaches. Notably, PA-VDM is model-agnostic and can be integrated with existing architectures like DiT or U-Net without architectural modifications. During training, PA-VDM fine-tunes pre-trained diffusion models with the new noise schedule to learn smoother scene transitions; at inference, it denoises frames progressively, generating long-form videos with improved consistency and visual fidelity.

While CausVid, ART-V, and PA-VDM focus on single-frame or short-interval autoregression, Henschel et al. [32]'s **StreamingT2V** pushes this further by generating videos in chunks of 8–16 frames in the latent space of a pretrained Vector Quantized Variational Autoencoder (VQ-VAE) variant, specifically a VQ-GAN [88], which compresses video frames into latent codes for efficient spatiotemporal diffusion. Training is conducted on short text-video datasets with diverse scenes, leveraging pretrained short-video generators (e.g., Modelscope) for initialization. Temporal coherence is achieved based on two key mechanisms: a Conditional Attention Module (CAM), which maintains short-term consistency by conditioning each video chunk on latent features from the preceding one, and an Appearance Preservation Module (APM), which retains high-level scene attributes from the initial chunk to prevent semantic drift over time. A randomized blending strategy further enables seamless extension of videos across chunk boundaries, supporting continuous generation without introducing artifacts or inconsistencies. As a result, StreamingT2V supports long-form video synthesis—ranging from 80 to over 1,200 frames (2+ minutes) at 720×720 resolution—with smooth transitions and consistent visual quality.

Focusing on compositionality and multi-object dynamics in long videos, Tian et al. [58] proposed **VideoTetris**, a diffusion-based framework combined with an auto-regressive architecture based on ControlNet to model complex spatio-temporal semantics by manipulating attention maps within the denoising network to track and distinguish multiple objects over time, enabling fine-grained control of their appearance, positioning, and interactions. A reference frame attention module improves temporal coherence by conditioning each frame on prior ones, generating coherent 10-second to 2-minute videos from either static or evolving prompts with dynamic object configurations. A motion-aware preprocessing pipeline further teaches the model to align complex textual cues with object dynamics.

Meanwhile, Qin et al. [50] proposed **xGen-VideoSyn-1**, a hybrid model for long-form video generation (over 14 seconds at 720p resolution) that combines latent diffusion architecture with a Video VAE (VidVAE) for spatial-temporal compression. The VidVAE extends the traditional VAE to 3D, enabling compression of video data both spatially and temporally, which significantly reduces the length of visual tokens and computational demands for generating long-sequence videos. To further address temporal coherence and computational efficiency, the model employs token segmentation by

 doi:10.20944/preprints202512.1476.v1

dividing lengthy videos into overlapping short segments of tokens, which are independently processed and then merged, maintaining temporal consistency across frames. Training is conducted on 13 million curated video-text pairs, collected through an automated pipeline that incorporates text detection, motion estimation, aesthetic scoring, and dense captioning.

While the above approaches focus on architectural modifications, Dalal et al. [25] explored a different direction by introducing Test-Time Training (TTT) layers to extend the temporal capacity of T2V models. Transformer-based models often struggle with long sequences due to the quadratic cost of self-attention and limited expressiveness of static hidden states, leading to degraded coherence over time. To address this, TTT layers replace standard hidden states with learnable neural modules that are fine-tuned during inference. This dynamic adaptation enables richer internal representations, improving temporal consistency and narrative complexity without requiring retraining on long videos. As a proof of concept, the authors curate a 7-hour dataset of annotated *Tom and Jerry* cartoons and initialize their model from a pre-trained diffusion Transformer (CogVideo-X 5B), originally limited to 3-second clips. By integrating TTT layers and restricting self-attention to 3-second segments, the extended model generates coherent one-minute videos (~63 seconds) at 16 FPS and around 256×256 resolution. Notably, the training proceeds in progressive stages, beginning with style transfer finetuning at 3-second lengths, then extending to 9, 18, 30, and 63 seconds to incrementally expand temporal context.

Balancing temporal coherence with visual fidelity, Bao et al. [19] introduced **Vidu**, a diffusion-based generation model using a U-shaped Vision Transformer (U-ViT) backbone combined with a video autoencoder. First, the autoencoder compresses raw video into a compact latent space. Then, within a U-Net–style encoder–decoder enhanced by skip connections, the U-ViT tokenizes these latents into 3D spatiotemporal patches. Self-attention across those patches captures long-range dependencies in both space and time. Finally, Vidu's diffusion process iteratively denoises the latents conditioned on text prompts to generate temporally coherent and high-resolution videos up to 1080p and 16 seconds in length in a single inference pass.

To address the limited accessibility of Sora and to democratize T2V research, Lin et al. [41] introduced **Open-Sora Plan**, an open framework designed for long-form, Sora-like video synthesis of up to 16 seconds at 720p resolution. Open-Sora Plan leverages a modular architecture that combines diffusion modeling, variational autoencoding, and transformer-based denoisers. Specifically, the framework integrates (i) a Wavelet-Flow Variational Autoencoder (VAE) for compressing videos into efficient, multi-scale latent representations, enabling the generation of high-resolution, temporally consistent video content. This is followed by (ii) a 3D full attention transformer denoiser utilizing Skiparse (sparse skip) Attention, which captures complex spatiotemporal dependencies across frames, and (iii) a conditional control module supporting multimodal inputs, such as textual prompts and reference images, for flexible and controllable generation. Open-Sora is trained via a progressive multi-stage strategy: it starts with image-pretrained weights and sparse attention to initialize spatiotemporal modeling, followed by joint training on both images and raw video at varying resolutions and durations, and finally fine-tuning on a cleaned Panda-70M subset. To ensure stable training, the framework incorporates adaptive gradient clipping and a min-max token strategy that balances resolution-duration buckets.

Lastly, Yang et al. [72] proposed **CogVideoX**, which replaces per-frame 2D VAEs by a 3D VAE that compresses video data jointly across spatial and temporal dimensions. This 3D structure preserves temporal continuity, reduces flicker, and enhances overall visual smoothness, while maintaining causal information flow to support long-range temporal modeling and improved reconstruction fidelity. To effectively bridge text and video modalities, CogVideoX introduces the Expert Transformer with Expert Adaptive LayerNorm (AdaLN). This architecture incorporates modality-specific normalization layers within a shared transformer backbone, enabling better processing for text inputs (e.g., from a T5 encoder) and video latents (from the 3D VAE), improving semantic alignment and content coherence. CogVideoX also leverages progressive training combined with multi-resolution frame packing, allowing the model to generate extended video sequences—up to 10 seconds at 16 FPS and 768×1360 resolution—while preserving visual coherence and dynamic diversity. Table 2 provides an overview of the studies mentioned above.

 doi:10.20944/preprints202512.1476.v1

Table 2. Summary of T2V models mentionned in the long-form video generation category.

Reference	Model Architecture	Methods	Training Strategy	Training Dataset	Project Code
Yin et al.	Transformer-based diffusion model	Bidirectional to autoregressive transformer adaptation + Distribution matching distillation + Asymmetric distillation + KV caching	Teacher-student distillation with ODE-based initialization + Asymmetric supervision + Training on short clips	UCF-101 (~13,000 clips)	https://github.com/tianweiy/CausVid
Weng et al.	Pretrained image diffusion model (Stable Diffusion 2.1) + Lightweight T2I-Adapter + Masked diffusion mask prediction	Autoregressive generation + Masked diffusion to reduce drifting	Learn simple continual motions + Noise augmentation + Anchored conditioning on initial frame	WebVid-10M	https://github.com/WarranWeng/ART.V
Xie et al.	Transformer-based latent diffusion model	Progressive noise assignment to latent frames + Autoregressive video denoising with overlapping attention windows	Autoregressive training with progressively increasing noise levels on latent frames	Large-scale, filtered datasets with ~1 million videos and 2.3 billion images (unnamed)	https://github.com/desaixie/pa_vdm
Henschel et al.	Autoregressive diffusion with CAM for short-term and APM for long-term appearance consistency + Video enhancer	Autoregressive diffusion with CAM and APM + Randomized blending for video enhancement	Initialization (pretrained model) + Autoregressive chunk generation + Streaming refinement with enhancer	Large-scale video collection from public sources, resized to 720×720 (unnamed)	https://github.com/Picsart-AI-Research/StreamingT2V
Tian et al.	Spatio-temporal compositional diffusion + ControlNet-based auto-regressive diffusion + Reference frame attention	Compositional video generation with spatial-temporal region modeling + Consistency regularization + Enhanced data preprocessing	Train auto-regressive model on enhanced video-text dataset	Filtered Panda-70M	https://github.com/YangLing0818/VideoTetris

Continued on next page

 doi:10.20944/preprints202512.1476.v1

Reference	Model Architecture	Methods	Training Strategy	Training Dataset	Project Code
Qin et al.	Video Variational Autoencoder (VidVAE) + Diffusion Transformer (DiT) with spatial & temporal attention	Latent Diffusion Model with Video VAE compression + Divide-and-merge strategy + Spatial-temporal self-attention	Progressive three-stage training at video resolutions of 240p, 480p, and 720p	13M+ curated video-text pairs (unnamed)	Not released
Dalal et al.	Pretrained Diffusion Transformer (e.g., CogVideo-X) augmented with adaptive TTT layers	Test-Time Training (TTT) layers updating during inference + Segmented video generation + Gating between local attention and TTT layers	Progressive fine-tuning on segmented cartoon videos from 3-second clips up to 63-second concatenations	Curated ~7 hours of Tom & Jerry videos + Human-generated storyboards	https://github.com/test-time-training/ttt-video-dit
Bao et al.	Diffusion model + U-ViT transformer backbone + Video autoencoder for compression	Diffusion model with U-ViT backbone + Video autoencoder + Transformer processes 3D patches + Re-captioning text	Train video captioner to auto-annotate video-text pairs + Multi-length video training	Large-scale text-video datasets with auto-generated captions (unnamed)	https://www.shengshu.com/en
Lin et al.	Diffusion Transformer (DiT) based STDiT; PixArt-α pretrained T2I backbone + spatial-temporal attention + pretrained spatial VAE	Wavelet-Flow Variational Autoencoder (VAE) + Joint Image-Video Skiparse Denoise + Conditional controller modules	Large-scale video pretraining on ~70M video clips, image pretraining from large image datasets, image-to-video finetuning	Curated ~70M video dataset from public sources; resized 256×256	https://github.com/PKU-YuanGroup/Open-Sora-Plan
Yang et al.	Transformer-based diffusion with 3D full attention + 3D VAE encoder-decoder + Expert Transformer with adaptive LayerNorm + 3D rotary positional embeddings	Latent diffusion denoising conditioned on expert fused text-video embeddings + LLaMA2 for video caption generation	Progressive training with mask modeling and long sequence modeling, separate VAE training	~35M curated video-text pairs + 2B images from LAION-5B and COYO-700M	https://github.com/THUDM/CogVideo

 doi:10.20944/preprints202512.1476.v1

b) User preference & human alignment

Addressing the critical challenge of aligning text-to-video (T2V) models with diverse human preferences, several works have focused on integrating user feedback and preference signals into video generation. For instance, Liu et al. [42] proposed **VideoDPO**, a pioneering pipeline that adapts Direct Preference Optimization (DPO)—originally developed for language and image generation—to video diffusion models. Pretrained T2V systems often struggle to balance visual quality and semantic fidelity across varied user expectations. VideoDPO addresses this gap through preference-driven fine-tuning tailored for video generation. At the core of VideoDPO is OmniScore, a unified metric that jointly captures visual fidelity and semantic alignment by combining intra-frame quality (clarity, aesthetics), inter-frame coherence (temporal consistency, motion stability), and text-video relevance using vision-language models. To generate training data, VideoDPO creates multiple video outputs per prompt using a base diffusion model, ranks them automatically via OmniScore, and forms preference pairs based on score differences—giving greater weight to pairs with larger gaps. The resulting re-weighted DPO loss fine-tunes the backbone model, improving alignment with user preferences across both visual and semantic dimensions.

Building on this line of research, Yin et al. [75] introduced **Step-Video-T2V**, a 30-billion-parameter T2V foundation model capable of generating videos up to 204 frames long. It employs a Deep Compression Video Variational Autoencoder (Video-VAE), achieving 16×16 spatial and $8\times$ temporal compression without sacrificing visual fidelity. The encoder utilizes causal Res3D blocks and a convolution-attention MidBlock, while the decoder symmetrically reconstructs frames from the compressed latent representation. A dual-path structure with causal 3D convolutions preserves high-frequency details while effectively compressing structural information. Generation is guided by a 48-layer Diffusion Transformer (DiT) with full 3D attention—48 heads per layer—trained using Flow Matching loss to improve denoising stability and spatiotemporal consistency. Bilingual generation in English and Chinese is enabled by dual pretrained multilingual text encoders conditioning the diffusion process. Finally, human-aligned fine-tuning via Video-based DPO enhances realism and temporal coherence by reducing artifacts based on preference signals. For evaluation, the authors propose Step-Video-T2V-Eval, a benchmark of 128 video prompts across 11 categories, enabling comparison against state-of-the-art open-source and commercial T2V systems.

In parallel, Wang et al. [59] proposed **TF-T2V**, a framework that improves semantic grounding and temporal coherence by decoupling text understanding from motion modeling. To overcome the scarcity and expense of labeled video-text datasets, TF-T2V leverages large-scale unlabeled videos from sources like YouTube to learn motion dynamics, while maintaining semantic alignment using limited labeled data. The model uses a dual-branch architecture built on a 3D-UNet diffusion model: a content branch for spatial generation conditioned on text, and a motion branch trained on unlabeled videos. Joint training with shared weights ensures motion coherence, and separating text decoding from motion modeling allows for exploiting text-free video data without relying on textual annotations. At the same time, it enhances text-conditioned spatial generation by incorporating high-quality image-text datasets such as LAION-5B.

c) physics-aware & reasoning-grounded generation

Traditional T2V models often struggle to simulate abstract physical laws, leading to implausible or inconsistent dynamics. To address this, Wang et al. [65] proposed **WISA**, a T2V generation framework that embeds interpretable physical knowledge into video generation for improved realism and consistency. WISA decomposes physical understanding into three components: (1) textual descriptions of expected physical behavior, (2) qualitative categories covering 17 physical phenomena across dynamics, thermodynamics, and optics, and (3) quantitative properties (e.g., density, temperature, refractive index). Its key innovation is the Mixture-of-Physical-Experts Attention (MoPA)—a multi-head attention mechanism in which each head specializes in a distinct physical category. A physical classifier identifies which phenomena are present in the input and selectively activates the relevant

 doi:10.20944/preprints202512.1476.v1

experts. Additionally, Adaptive Layer Normalization (AdaLN) is used to embed continuous physical values, allowing fine-grained modulation of video generation based on real-world measurements. To support training and evaluation, the authors release **WISA-32K**, a curated dataset of 32,000 videos labeled with physical categories and properties to support physics-aware generation.

Building on the idea of physics-grounded synthesis for natural phenomena, Yuan et al. [77] proposed **MagicTime** for metamorphic time-lapse video generation that learns real-world physical transformations from time-lapse data. It introduces the MagicAdapter module that decouples spatial and temporal learning, enabling pretrained T2V models to better capture long-term physical variability of metamorphic phenomena, while retaining their general video synthesis capabilities. To improve temporal fidelity, MagicTime incorporates a Dynamic Frames Extraction strategy, which prioritizes frames with significant transitions to focus learning on key metamorphic moments, while a custom Magic Text-Encoder enhances alignment by distinguishing metamorphic-specific language. The model is trained on ChronoMagic, a curated dataset of 2,265 time-lapse videos with auto-generated captions focused on persistent, physically meaningful transformations. Cascade preprocessing and multi-view text fusion integrate diverse textual and dynamic perspectives during preprocessing to further improve video quality and prompt understanding.

Shifting from domain-specific physics to general realism, Qing et al. [51] proposed **HiGen**, a diffusion-based framework that hierarchically decouples spatial and temporal components of video synthesis to improve realism and motion stability. By disentangling structure and content, HiGen effectively addresses the challenges posed by the intricate interplay between spatial details and motion dynamics. At the structure level, training is divided into two sequential stages using a unified denoiser. First, it performs spatial reasoning to produce coherent static priors conditioned on text input. These priors then guide temporal reasoning, which generates smooth and temporally coherent motion. At the content level, the model extracts two subtle cues from input video content during training—a motion-related cue that captures dynamic changes and appearance-related cue that reflects variations in visual style or content. These cues improve the model's ability to handle spatial and temporal variations independently while maintaining overall coherence.

While physics-based models focus on capturing natural phenomena through learned physical transformations and domain-specific dynamics, some works integrate large language models (LLMs) to provide high-level reasoning, planning, and semantic structure for video generation. For instance, Fei et al. [26] introduced **Dysen-VDM**, a diffusion model that leverages LLMs for enhanced scene understanding and action planning. It incorporates a Dynamic Scene Manager (Dysen) module inspired by human cognitive intuition, which imposes structured temporal awareness during video generation. Dysen extracts and orders actions from text, builds a dynamic scene graph (DSG) to capture spatiotemporal relations, and enriches it using in-context LLMs (e.g., ChatGPT). The enriched DSG is encoded via a recurrent graph Transformer and injected into the backbone latent VDM. The training follows three stages: (i) pre-train the backbone VDM with an autoencoder on the WebVid dataset to learn initial video representations, (ii) continue pre-training for text-conditioned video generation on WebVid while integrating the recurrent graph Transformer encoder for DSG processing, and (iii) fine-tune the full Dysen-VDM with Dysen, using reinforcement learning to optimize in-context learning for more coherent scene graphs.

Similarly, Lv et al. [44] proposed **GPT4Motion**, a training-free framework designed to improve T2V generation by combining the planning capabilities of LLMs like GPT-4, the physical simulation power of Blender, and the visual quality of text-to-image diffusion models such as Stable Diffusion. The pipeline begins with GPT-4, which translates a user's textual prompt into an executable Python script using Blender's API to construct scenes and simulate physics-driven dynamics, including object collisions, cloth deformation, and fluid motion. Blender then runs the simulation and produces structured scene data (e.g., depth maps, segmentation masks, edge maps) that encode the physical evolution of the scene. These temporally consistent representations are passed to a text-to-image diffusion model,

 doi:10.20944/preprints202512.1476.v1

which, conditioned on both the original prompt and simulated scene priors, synthesizes high-quality video frames.

Beyond guiding physical realism, LLMs can also act as semantic planners. For instance, Zhou et al. [85] proposed **HiTVideo**, a T2V framework that improves text-video alignment through a hierarchical video tokenizer built on a 3D causal VAE. It encodes video content into multiple discrete codebooks, where high-level tokens capture global semantics (e.g., scene layout, object composition, motion), while low-level ones retain fine-grained spatiotemporal textures necessary for high-fidelity reconstruction. This hierarchical design reduces bits-per-pixel by approximately 70% compared to standard tokenizers, with minimal visual quality loss. For generation, an autoregressive LLM (e.g., LLaMA-3B), conditioned on a frozen text encoder (e.g., Flan-T5-XL), produces tokens in a coarse-to-fine manner, beginning with high-level semantic tokens and progressively refining output with lower-level detail tokens. Positional encodings along spatial and temporal axes enable coherent video generation (64+ frames, ~8 seconds). Hierarchical tokenization shortens LLM input and improves semantic alignment. Additional features include dynamic encoding (adjusting compression by scene complexity) and masked decoding (for efficient selective token prediction). Training follows a two-stage pipeline: first, pretraining the VAE on unlabeled videos; second, training the LLM to autoregressively generate video tokens from text prompts and prior tokens, with masked modeling and adaptive positional encoding to handle transitions and variable lengths.

Finally, Tan et al. [57] proposed **Mimir**, an end-to-end training framework that significantly improves T2V diffusion models by combining LLMs for advanced semantic reasoning alongside conventional text encoders. Mimir uses a dual-branch architecture: a standard text encoder (e.g., T5) captures local syntax and structure, while a decoder-only LLM (e.g., Phi-3.5) models global semantics and contextual imagination. A token fuser module then aligns and integrates these heterogeneous outputs. This fused representation conditions a latent diffusion model, enabling semantically coherent and syntactically grounded video generation within a compact latent space. The dual-branch encoders, token fuser, and diffusion backbone are trained jointly on a curated large-scale dataset of ~500,000 video clips, each averaging 10 seconds in length. An overview of the papers discussed in categories b) and c) is presented in Table 5.

 doi:10.20944/preprints202512.1476.v1

Table 3. Summary of T2V models mentioned in categories b) and c).

Reference	Model Architecture	Methods	Training Strategy	Training Dataset	Project Code
Liu et al.	Post-training adaptation of existing diffusion models; no architecture change	Direct Preference Optimization (DPO) + Omni-preference alignment + VideoDPO Loss	Multi-stage: T2V pretraining → DPO fine-tuning with re-weighted preference pairs	VidProM dataset with 10,000 human prompts + millions of videos for preference alignment	https://github.com/CIntellifusion/VideoDPO
Yin et al.	Video-VAE with causal Res3D blocks + DiT with 3D attention + bilingual text encoders; + Video-DPO module	Latent video diffusion, Flow Matching training, multilingual text encoding, preference-based Direct Preference Optimization (Video-DPO)	Multi-stage training: Video-VAE pretraining → DiT diffusion model training → DPO finetuning	Large-scale, filtered video corpus (unnamed)	https://github.com/stepfun-ai/Step-Video-T2V
Wang et al.	Two-branch diffusion model: content branch (text-conditioned), motion branch (image-conditioned) with shared weights	Disentangled content-motion learning, temporal coherence loss	Joint optimization on video-text and text-free video data; semi-supervised approach	WebVid10M video-text pairs + large-scale unlabeled videos from YouTube	https://github.com/ali-vilab/Vgen
Wang et al.	Enhanced T2V model with physics-aware modules	Decomposition of physical principles into textual, qualitative, and quantitative components; MoPA attention; Physical classifier	Joint training with generative loss and physical classifier supervision	WISA-32K: 32,000 videos on 17 physical laws (dynamics, thermodynamics, optics)	https://github.com/360CVGroup/WISA
Yuan et al.	Diffusion Transformer (DiT) backbone with MagicAdapter modules in temporal layers, enhanced text encoder	MagicAdapter for physics encoding; Dynamic Frame Extraction for adaptive sampling + Magic Text-Encoder for prompt understanding + auto-captioning for annotation	Fine-tuning pretrained T2V models on metamorphic time-lapse videos	Curated ChronoMagic dataset (~2,265 annotated metamorphic time-lapse videos)	https://github.com/PKU-YuanGroup/MagicTime

Continued on next page

 doi:10.20944/preprints202512.1476.v1

Reference	Model Architecture	Methods	Training Strategy	Training Dataset	Project Code
Qing et al.	Diffusion model with unified denoiser for spatial-temporal reasoning + motion and appearance cue extraction modules	Decoupled spatial-temporal diffusion + hierarchical conditioning on motion and appearance variations	Two-step training: spatial prior from text → temporal motion from spatial priors	MSR-VTT (10,000 video-text pairs)	https://github.com/ali-vilab/Vgen
Fei et al.	Latent video diffusion model + Dysen module (action extractor, DSG constructor, LLM-based enrichment, graph Transformer encoder)	Dynamic Scene Graphs (DSGs) + LLMs (ChatGPT) + Recurrent Graph Transformer (RGT)	Pre-training on large-scale video-text datasets → Fine-tuning Dysen module and RGT	3M WebVid (pre-training) + UCF-101, MSR-VTT, ActivityNet (fine-tuning)	https://github.com/scofield7419/Dysen
Lv et al.	Pipeline of GPT-4 (planner) + Blender physics engine (simulator) + ControlNet-augmented Stable Diffusion	GPT-4 generates Blender scripts for simulation + Blender simulates motion + Stable Diffusion generates frames	Training-free framework; uses pretrained GPT-4 and Stable Diffusion; no fine-tuning	None	https://github.com/jiaxilv/GPT4Motion
Zhou et al.	3D causal VAE with hierarchical discrete token layers + LLaMA 3B for token generation conditioned on frozen Flan-T5-XL embeddings	Hierarchical autoregressive token generation from coarse semantic to fine visual detail	Two-stage training: Training of the 3D causal VAE → Training of LLM autoregressively with text conditioning	Pexels Videos dataset	Not released
Tan et al.	Latent diffusion backbone + Dual-text encoding (ViT-style encoder + decoder-only LLM) + Token Fuser	Token fusion to harmonize encoder and LLM embeddings + Semantic stabilization	Joint training on large video-text datasets with supervised latent denoising conditioned on fused semantic embeddings	Large-scale video-text corpora (unnamed)	https://lucaria-academy.github.io/Mimir

 doi:10.20944/preprints202512.1476.v1

d) Hybrid & factorized generation

Unlike end-to-end generative models, some approaches leverage hybrid and factorized generation pipelines that decompose the generation process into modular stages or explicitly separate spatial and temporal modeling components, thereby improving synthesis quality and efficiency while reducing computational overhead.

For instance, Girdhar et al. [28] introduced **Emu Video**, a T2V model that decomposes video generation into two stages: (1) generating a high-quality image from a text prompt using a frozen text-to-image latent diffusion model, and (2) synthesizing a video conditioned on both the text and the generated image. This factorized approach simplifies generation compared to end-to-end or cascaded models, with the static image capturing scene composition and visual details, while the video diffusion transformer animates the scene over time. Emu Video employs tailored noise schedules and a multi-phase training strategy to achieve stable, high-resolution (512×512) outputs with smooth temporal dynamics, avoiding the complexity and overhead of multi-stage diffusion pipelines.

Focusing on disentangled modeling, Chen et al. [22] presented **VideoCrafter2**, a method designed to improve the visual fidelity when training on low-quality, large-scale datasets such as WebVid-10M. The core idea is to disentangle spatial and temporal learning: the architecture employs factorized 3D U-Net temporal modules that separately model temporal dynamics alongside pretrained spatial modules responsible for appearance and semantic grounding. The full model is first trained on video datasets (WebVid-10M, LAION-COCO 600M [89]) to capture motion and semantic grounding, after which only the spatial modules are fine-tuned on high-quality synthetic images (e.g., Midjourney, JDB) to enhance visual fidelity while preserving learned motion dynamics. Additional refinements include frame rate (FPS) conditioning for controllable temporal resolution and the use of LoRA for parameter-efficient fine-tuning.

Similarly, Menapace et al. [46] introduced **Snap Video**, a video-first transformer architecture optimized for speed and scalability. Rather than extending 2D image-based U-Nets with temporal layers—resulting in heavy computational costs—Snap Video employs a factorized spatiotemporal architecture that jointly models spatial and temporal features in a compressed latent space using a scalable transformer. The approach adapts the Efficient Diffusion Model (EDM) framework to handle the spatial and temporal redundancies common in video, enabling more natural and efficient generation. Training proceeds in two stages: an initial 550k-step training phase, followed by fine-tuning for 370k steps on high-resolution videos. Processing spatiotemporal information as a unified 1D latent vector achieves a 3.31× speedup in training and 4.5× faster inference compared to U-Net baselines.

Expanding on multi-stage generation pipelines, Wang et al. [60] proposed **LaVie**, a T2V generation framework that builds on Stable Diffusion to generate high-quality, temporally coherent videos. It uses a three-stage cascaded latent diffusion architecture: (1) generates low-resolution short clips conditioned on text, (2) synthesizes smooth intermediate frames to increase frame rate, and (3) upsamples the result to high-definition video. The pipeline starts with a base T2V model—initialized from a pretrained Stable Diffusion checkpoint and extended with temporal modules (pseudo-3D convolutions and spatiotemporal self-attention)—to produce semantically aligned low-resolution clips. A temporal interpolation model then increases frame rate (4×) by synthesizing intermediate frames for smoother motion. Finally, a video super-resolution module (fine-tuned from image SR models) upsamples the interpolated videos (e.g., to 1280×2048), improving spatial detail while preserving temporal continuity. Temporal coherence is achieved through self-attention layers with rotary positional encoding. LaVie is jointly fine-tuned on image and video data to maintain visual diversity and realism. To support training, the authors introduce Vimeo25M, a large-scale dataset of 25M diverse, high-quality text-video pairs.

Finally, Zhang et al. [83] proposed **Show-1**, a hybrid T2V generation framework that integrates both pixel-based and latent-based VDMs for improved semantic alignment and computational efficiency. It employs a two-stage, coarse-to-fine generation pipeline. In the first stage, a pixel-based VDM is trained to generate low-resolution videos (e.g., 64×40), leveraging image-pretrained weights (e.g., from LAION) for strong semantic grounding and coherent motion generation. In the second

 doi:10.20944/preprints202512.1476.v1

stage, a latent diffusion model serves as a super-resolution expert that refines and upscales outputs to higher resolutions (e.g., 256×160), enhancing visual fidelity while preserving the content structure and alignment achieved in the earlier stage. Both stages share a pretrained T5 text encoder for consistent semantic guidance, while motion customization is achieved by fine-tuning only the temporal attention layers. The first stage employs full 3D spatiotemporal attention for low-res video generation, and the second applies 1D spatial upscaling and temporal interpolation for frame-wise super-resolution and smooth motion.

 doi:10.20944/preprints202512.1476.v1

Table 4. Summary of T2V models mentioned in hybrid & factorized generation.

Reference	Model Architecture	Methods	Training Strategy	Training Dataset	Project Code
Girdhar et al.	Two-stage factorized generation: frozen T2I image + latent video diffusion with temporal layers & image conditioning	Factorized conditioning + tuned noise schedules + classifier-free guidance with separate image/text weights	Frozen T2I init + multi-stage training: 256px image-conditioned → 512px zero terminal-SNR → high-motion subset fine-tune (1.6K clips)	34M licensed text-video pairs (unnamed)	https://emu-video.metademolab.com
Chen et al.	Stable Diffusion backbone + factorized 3D U-Net temporal modules + separate spatial & temporal modules	Disentangled spatial/temporal training + partial temporal tuning + LoRA fine-tuning + frame rate conditioning	Temporal modules trained on low-quality videos → spatial modules trained on high-quality images → separate fine-tuning	WebVid-10M (low-quality videos) + JDB (Midjourney-synthesized high-quality images)	https://github.com/AILab-CVC/VideoCrafter
Menapace et al.	Spatiotemporal transformer (FIT) + extended EDM diffusion for video	Joint modeling of spatial and temporal redundancies	Two-stage training: pre-training on lower-res videos → fine-tuning on high-res videos	Large-scale video-text datasets (unnamed)	https://snap-research.github.io/snapvideo
Wang et al.	Cascaded latent diffusion models + temporal self-attention + rotary positional encoding	Latent diffusion + temporal self-attention for frame coherence + temporal interpolation in latent space	Joint image-video fine-tuning + cascaded training	Curated Vimeo25M dataset with 25 million text-video pairs	https://github.com/Vchitect/LaVie
Zhang et al.	Two-stage hybrid: pixel-based diffusion for low-res generation + latent diffusion for high-res upscaling	Hybrid pixel-latent diffusion pipeline + expert translation module for super-resolution	Multi-stage training (keyframe generation → frame interpolation → SR → expert fine-tuning)	WebVid-10M	https://github.com/showlab/Show-1

 doi:10.20944/preprints202512.1476.v1

4.1.2. Controllable T2V Generation

Unlike fully automatic text-to-video generation, controllable video generation focuses on synthesizing videos in which specific aspects of the content—such as object appearance, spatial layout, motion trajectory, temporal dynamics, or style—can be manipulated by the user. This category covers many control modalities, including trajectory & motion trajectory control, camera & pose control, entity- & object-level Control, multi-prompt & temporal editing, and frequency- & identity control. Figure 8 illustrates the works from the selected literature, organized into these subcategories, which are detailed in the following sections.

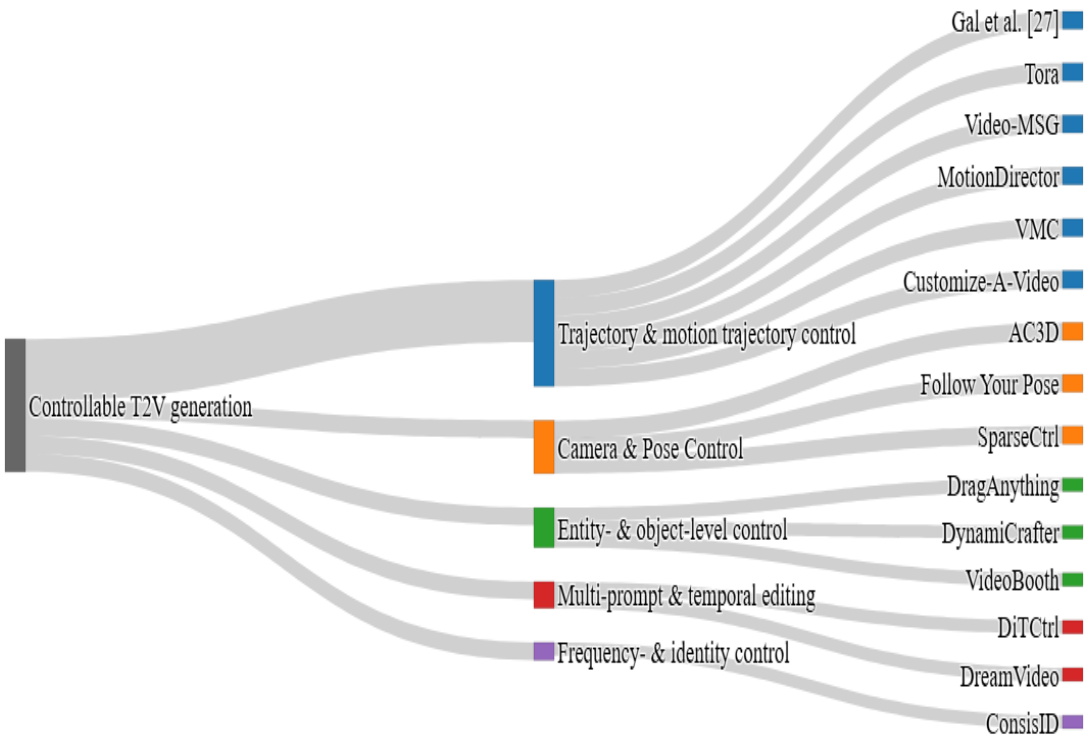

Figure 8. Schematic overview of T2V model classified within the controllable T2V generation category.

a) Trajectory & motion trajectory control

In T2V generation, it refers to methods that allow users to define the spatial and temporal movements of objects, camera viewpoints, and scene elements. Rather than relying solely on text prompts, these approaches incorporate user-defined motion specifications—such as predefined paths, keyframes, or movement constraints—to guide the generation process.

One approach provides *indirect* motion control by extracting motion priors from pretrained models. For example, Gal et al. [27], who introduced a method to animate static sketches using pretrained T2V diffusion models guided by text prompts. Their approach distills motion priors from large T2V models to produce short, semantically meaningful vector-based animations. Instead of training a dedicated model, their approach distills motion priors from large T2V models to produce short, semantically meaningful vector-based animations. Motion is decomposed into local stroke deformations and global affine transformations, ensuring both fine detail and structural coherence. The core innovation lies in a score-distillation optimization scheme that guides the stroke deformation and transformation parameters using the pretrained diffusion model's learned video distribution. This optimization evaluates how semantically meaningful and natural a given motion is relative to the input

 doi:10.20944/preprints202512.1476.v1

text prompt. By minimizing the score-distillation loss, the method generates smooth, semantically consistent animations without requiring costly or extensive additional training.

By contrast, some works pursue *direct* trajectory conditioning so users can draw or specify exact paths. For instance, Zhang et al. [82] introduced **Tora**, the first trajectory-oriented Diffusion Transformer (DiT) framework, enabling controllable video generation conditioned on text, image, and explicit 2D motion trajectories for precise user guidance. At its core, the Trajectory Extractor (TE), which encodes 2D trajectories—represented as sequences of points—into a compact latent forms by converting them into RGB flow-like maps using flow visualization and Gaussian smoothing, then encoding them with a 3D VAE into latent motion patches capturing rich dynamics. The backbone, Spatial-Temporal Diffusion Transformer (ST-DiT), alternates between Spatial DiT Blocks (spatial self-/cross-attention) and Temporal DiT Blocks (temporal self-attention), capturing long-range spatial and temporal dependencies. To incorporate trajectory guidance, the Motion-Guidance Fuser (MGF) injects the TE's motion patches into the transformer layers, adaptively normalizing their activations based on motion embeddings to ensure path adherence. Training proceeds in two stages: pretraining on dense optical flow datasets to learn motion, followed by fine-tuning on sparse trajectory annotations to enable interactive user control.

Bridging planner-style sketches and direct trajectory conditioning, Li et al. [39] proposed **Video-MSG**, a training-free guidance framework that enhances T2V diffusion models' ability to follow complex text prompts involving spatial layouts and object trajectories. Without modifying the underlying T2V model or incurring significant memory or inference costs, Video-MSG enables precise and controllable video generation. The framework operates in three stages. First, it performs multimodal planning to generate a Video Sketch—a sequence of draft frames encoding background, foreground, and object motion as a fine-grained spatio-temporal blueprint. This sketch is derived from multimodal inputs, including text, detected objects, and scene segmentation using pretrained object detectors, instance segmentation, and a multimodal large language model (MLLM) for spatial and temporal planning. Second, in structured noise initialization, the Video Sketch is used to initialize the latent noise input to the T2V model, injecting spatially and temporally coherent priors that guide generation toward faithful layouts and trajectories while improving alignment with the text prompt. Finally, through guided diffusion involving noise inversion and denoising, the model synthesizes videos that closely match the intended layout and dynamics specified in the sketch and text prompt.

A different line of research focuses on *motion transfer*: extracting motion from a reference clip and applying it to new subjects while preserving appearance diversity. Zhao et al. [84] proposed **MotionDirector**, a framework that disentangles motion from appearance and frames motion customization as adapting a pretrained T2V model to reproduce reference motions (e.g., a car navigating a path) independently of subject appearance. It introduces a dual-path LoRA architecture: the spatial path captures appearance features by training spatial LoRAs on randomly sampled single frames, ensuring independence from motion cues; the temporal path reuses these spatial LoRAs for consistency while adding temporal LoRAs trained on frame sequences to model motion dynamics. A key contribution is the appearance-debiased temporal loss, which mitigates appearance interference during temporal learning, enabling the temporal LoRAs to improve motion representation. This separation allows motion patterns to generalize across subjects. By updating only low-rank adapters and freezing the base model, it achieves efficient adaptation without full retraining.

Similarly, Jeong et al. [34] introduced **VMC** (Video Motion Customization), a framework for customizing T2V diffusion models to generate videos with user-specified motion patterns. VMC applies one-shot tuning to the temporal attention layers of pre-trained models, enabling the integration of reference motion without disrupting scene or appearance generation. It introduces a motion distillation loss that encodes motion using residual latent-frame differences, guiding the model to reproduce smooth, low-frequency motion while avoiding high-frequency artifacts. To ensure motion transfer is decoupled from appearance, VMC converts prompts into appearance-invariant forms by removing or

 doi:10.20944/preprints202512.1476.v1

neutralizing specific background or subject details, enabling consistent motion reproduction across varied scenes.

Lastly, Ren et al. [54] proposed **Customize-A-Video**, a framework for one-shot motion customization in T2V diffusion models. It tackles the challenge of transferring specific motion patterns from a single reference video to novel subjects or scenes with varying spatiotemporal properties. The core innovation lies in one-shot motion learning, where motion dynamics are captured from a single video example without requiring large-scale motion-labeled datasets. The framework fine-tunes a pretrained 3D U-Net–based diffusion model—with spatial and temporal transformer blocks—using LoRA on temporal attention layers to capture fine-grained motion while preserving the model's generative capacity for content and appearance. Moreover, the authors introduce appearance absorbers, modules pretrained on image or video data to isolate and neutralize spatial appearance features from the reference video before motion adaptation. These absorbers help disentangle motion from appearance, enabling motion transfer across diverse contexts. The training follows a two-stage pipeline while keeping the pretrained T2V backbone frozen: first, appearance absorbers are trained; then, with appearance signals absorbed, temporal LoRAs are trained on temporal attention layers to learn motion representations. This staged approach ensures progressive motion learning and appearance control, with plug-and-play modules for flexible integration into existing T2V systems.

b) Camera & Pose Control

In T2V generation, it refers to methods that enable users to manipulate camera viewpoints and motion (e.g., position, rotation, focal length, and temporal trajectories) as well as human or character poses and their temporal trajectories, thereby enabling the creation of cinematic visual content.

For example, Bahmani et al. [18] introduced **AC3D**, a novel architecture designed to achieve precise and high-quality 3D camera control within video diffusion transformer models. Built on transformer-based video diffusion backbones—extending VD3D and VDiT architectures—AC3D processes video latent tokens alongside dedicated camera tokens. Camera pose information is encoded using Plücker coordinates and transformed into camera tokens by a fully convolutional encoder, ensuring spatial and channel compatibility with video tokens. These camera tokens are further processed by lightweight DiT-XS style transformer blocks before conditioning.

The authors start by analyzing camera motion from first principles, demonstrating that camera-induced movement primarily manifests in the low-frequency components of video sequences. Building on this insight, they propose redesigned pose conditioning schedules for both training and inference, which accelerate convergence while enhancing both visual and motion quality in generated videos. Linear probing reveals that camera pose information is predominantly encoded in the 30% of network layers. Consequently, AC3D restricts the injection of camera conditioning inputs to these early layers, reducing trainable parameters by 4× and improving visual fidelity by 10%. To better disentangle camera and scene motion, AC3D introduces a 20K-video dataset recorded with stationary cameras, enhancing the model's ability to generate natural, pose-conditioned videos.

Extending pose controllability to characters, Ma et al. [45] proposed **Follow Your Pose**, a two-stage framework for generating pose-controllable, text-editable character videos without requiring paired video-pose-caption data. The approach builds on a pretrained text-to-image (T2I) diffusion model backbone (e.g., Stable Diffusion) extended for video generation. In the first stage, the model learns pose-conditioned text-to-image (T2I) generation from keypoint-image pairs by training a zero-initialized convolutional encoder that injects pose information into the pretrained T2I pipeline while preserving its editing and compositional abilities. The second stage fine-tunes this model on large-scale pose-free videos by introducing learnable temporal self-attention and cross-frame attention blocks, enabling temporal dynamics and motion consistency without explicit pose labels.

Finally, Guo et al. [29] presented **SparseCtrl**, a plug-and-play framework that enhances controllability in T2V diffusion models via temporally sparse structural conditioning. Unlike traditional methods that require dense per-frame conditioning inputs (e.g., depth maps or sketches), SparseCtrl uses only a few condition frames scattered across the video, significantly reducing annotation overhead.

 doi:10.20944/preprints202512.1476.v1

It incorporates a modality-agnostic condition encoder (sketch/depth/RGB) with lightweight adapter heads, while keeping the pretrained T2V backbone frozen; only the encoder and adapters are optimized on WebVid-10M, avoiding costly full-model retraining. The encoder reuses frame-wise 2D layers from the backbone and adds temporal-aware transformer blocks to propagate sparse conditioning across frames. During training, random temporal sparse masks simulate different sparsity levels by varying the number and positions of conditioning frames, improving robustness at inference.

c) Entity- & object-level control

In T2V generation, it refers to methods that enable users to manipulate individual objects or entities within a video at a fine-grained semantic level—controlling appearance, motion, spatial placement, and interactions via signals beyond text prompts such as reference images and user-drawn trajectories.

For instance, Wu et al. [69] proposed **DragAnything**, a framework for controllable video generation that enables precise, intuitive motion manipulation at the entity level. Unlike prior approaches that rely on pixel-level dragging or sparse trajectories—which often fail to capture the semantics or structure of entire objects—DragAnything introduces a representation-driven method capable of controlling the motion of individual objects, background elements, or multiple entities simultaneously. The framework is built on top of the latent diffusion video generation model Stable Video Diffusion (SVD), which employs a 3D U-Net backbone to encode and decode video latent representations. The core innovation lies in its entity representation extraction: the framework leverages latent features from a foundational diffusion model, indexed via segmentation masks (e.g., from the Segment Anything Model (SAM)), to encode complete object-level embeddings. Users guide video generation by drawing motion trajectories, which are converted into 2D Gaussian-weighted signals focused on central pixels of the entity masks to emphasize key object regions. The conditional video denoising autoencoder then synthesizes frames conditioned on the initial video frame, the entity semantic embedding, and the Gaussian trajectory signal, producing videos reflecting the desired motion. Training employs annotated video segmentation benchmarks with object masks and tracking data. Motion trajectory supervision is generated using *Co-Tracker*, which produces object center trajectories over time. The model learns via supervised loss concentrated on masked regions to generate artifact-free, localized motion while preserving the rest of the video.

Where DragAnything emphasizes entity embeddings and trajectories, other work shows how motion priors from pretrained T2V models can animate arbitrary imagery. For instance, Xing et al. [71] presented **DynamiCrafter**, a framework for animating still images using motion priors from pretrained T2V diffusion models. Unlike traditional animation techniques often limited to specific domains (e.g., humans or natural scenes), DynamiCrafter generalizes across diverse visual content—including objects, animals, CGI, and art—while preserving fine visual details, supporting high-resolution generation (up to 576×1024) and sequences of up to 16 frames. It employs a dual-stream image injection paradigm. In the *semantic context stream*, the input image is projected into a semantic context space using a CLIP image encoder coupled with a learnable query transformer. This representation is integrated into the T2V diffusion model via cross-attention, allowing the model to understand the global structure and semantics of the scene in a way aligned with its pretrained feature space. The *visual detail stream* ensures visual detail guidance by concatenating the full-resolution input image with the model's initial latent noise to preserve low-level appearance. Training follows a three-stage process: (1) pretraining the context network—including the query Transformer—on a lightweight T2I model to encode semantics; (2) adapting it to the T2V backbone by jointly training with the spatial layers while freezing temporal layers to preserve motion priors; and (3) jointly fine-tuning the context network and spatial layers with the image–noise concatenation to improve appearance fidelity without degrading temporal dynamics.

Finally, Jiang et al. [35] introduced **VideoBooth**, a diffusion-based video generation framework that integrates both image and text prompts to offer precise, user-controlled video synthesis. VideoBooth preserves subject identity by encoding image prompts through a coarse-to-fine strategy: At the coarse level, a pretrained CLIP image encoder extracts high-level visual features from the image

 doi:10.20944/preprints202512.1476.v1

prompt, which are refined via multi-layer perceptrons trained to map them into the text embedding space. These coarse visual embeddings are fused with CLIP text embeddings to guide the overall semantics of the video. At the fine level, an attention injection module extracts multi-scale image features and injects them into the cross-frame attention layers of the video diffusion transformer, refining spatial details and thus temporal coherence throughout the video. The MLP-based visual embedding encoder and the attention injection module are jointly trained on WebVid-10M while keeping the pretrained T2V backbone frozen. At inference, VideoBooth operates feed-forward without finetuning.

d) Multi-prompt & temporal editing

In T2V generation, it refers to methods that enable coherent and semantically consistent video synthesis guided by multiple sequential or overlapping text prompts over time. Unlike the traditional single-prompt methods that generate videos from a static instruction, these approaches handle dynamic scenarios where the content, scene, or actions evolve with changing textual inputs.

For example, Cai et al. [20] proposed **DiTCtrl**, a training-free approach for generating long-form videos from multiple sequential text prompts within the Multi-Modal Diffusion Transformer (MM-DiT) framework—a transformer-based video diffusion backbone employing full 3D attention analogous to the cross/self-attention in UNet-style diffusion models. Unlike existing models that primarily handle single-prompt inputs and often produce disjointed results when given multiple prompts, DiTCtrl frames multi-prompt generation as a temporal video editing task—ensuring smooth transitions and consistent object motion across prompt boundaries without retraining. Its core innovation is a mask-guided attention control mechanism that modifies MM-DiT's 3D full attention to token subsets associated with each prompt, ensuring precise per-segment conditioning. To preserve semantic continuity across segments, a key–value sharing strategy propagates context between attention layers, maintaining consistent object identities and motion flows across prompt boundaries. To further enhance temporal coherence, a latent blending strategy merges overlapping latent video representations at segment transitions using position-dependent weights to prevent artifacts and abrupt visual changes.

In parallel, Wei et al. [66] presented **DreamVideo**, a personalized T2V framework that generates videos of custom subjects (from images) performing custom motions (from videos). Unlike prior approaches that focus exclusively on either subject or motion personalization, DreamVideo is the first to decouple subject identity and motion learning within a unified architecture, enabling flexible and composable video generation. Built on a pretrained video diffusion backbone—a U-Net with temporal attention and convolutional layers—DreamVideo introduces two lightweight adapters: an *identity adapter*, integrated primarily into the spatial cross-attention layers to incorporate fine-grained subject-specific features, and a *motion adapter*, inserted into temporal layers to capture motion dynamics. The method follows a two-stage decoupling strategy. In the *subject learning stage*, a textual identity embedding is first learned from a few static images via Textual Inversion to represent general subject characteristics. This embedding is then paired with a zero-initialized bottleneck identity adapter to capture detailed appearance features while keeping the pretrained diffusion weights frozen. In the *motion learning stage*, the motion adapter is trained on videos exemplifying the target motion pattern, with the identity embedding and adapter frozen to preserve appearance. During inference, the independently trained adapters are combined without retraining.

e) Frequency- & identity control

In T2V generation, it refers to methods designed to generate videos that preserve human identity—especially facial fidelity—across video frames by leveraging frequency-domain analysis. For instance, Yuan et al. [79] introduced **ConsisID**, a tuning-free T2V generation framework designed to address this challenge. The framework is built on a pretrained Diffusion Transformer (DiT) backbone, which replaces conventional U-Net architectures with high-capacity transformer blocks for video denoising. ConsisID decomposes facial identity into low- and high-frequency components corresponding to

 doi:10.20944/preprints202512.1476.v1

different identity representations. Low-frequency features, encoding global aspects such as face shape and structure, are extracted using a global facial feature module and injected into early transformer layers to stabilize identity. High-frequency features, encoding fine details like skin texture and subtle appearance cues, are derived from a local extractor and integrated into deeper layers to preserve appearance realism. A hierarchical training strategy adapts a pretrained video diffusion model by incorporating these frequency-specific features, transforming it into an identity-preserving T2V system (IPT2V) that ensures coherence and realism across generated frames.

 doi:10.20944/preprints202512.1476.v1

Table 5. Summary of T2V models mentioned in Controllable T2V generation.

Reference	Model Architecture	Methods	Training Strategy	Training Dataset	Project Code
Gal et al.	Lightweight network controlling sketch strokes + pretrained T2V motion prior with local deformation and global affine components	Score distillation sampling (SDS) loss + vector Bézier curve sketch representation	Optimization-based, no training or fine-tuning	None	https://github.com/yael-vinker/live_sketch
Zhang et al.	DiT backbone (OpenSora) with: Trajectory Extractor (3D VAE) + Spatial-Temporal DiT blocks + Motion-guidance Fuser	Trajectory-conditioned video generation + MGF hierarchical fusion + diffusion with text/visual conditions + alt. spatial-temporal attention	3D Training of 3D VAE on flow maps + Joint training of diffusion transformer and MGF on trajectory-annotated video-text data	630,000 videos from: Panda-70M + Mixkit + Pexels + Internal sources	https://github.com/alibaba/Tora
Li et al.	Pretrained T2V diffusion model + pipeline of multimodal LLM planner	Background planning, foreground layout & trajectory planning + structured noise init. + MLLM/vision models	No training or fine-tuning	None	https://github.com/jialuli-luka/Video-MSG
Zhao et al.	Pretrained 3D U-Net T2V diffusion backbone + dual-path spatial & temporal LoRA modules	Motion customization via decoupled LoRA tuning + appearance-debiased temporal loss + Temporal Attention Purification	Fine-tune spatial LoRAs on single frames + temporal LoRAs on multiple frames + backbone frozen	UCF Sports Action	https://github.com/showlab/MotionDirector
Jeong et al.	Pretrained cascaded VDM backbone + adapted temporal attention layers	Motion distillation via residual latent frame vectors + appearance-invariant prompt transformation	Parameter-efficient fine-tuning on temporal attention layers only + motion distillation loss + frozen backbone	Few, short videos	https://github.com/HyeonHo99/Video-Motion-Customization

Continued on next page

 doi:10.20944/preprints202512.1476.v1

Reference	Model Architecture	Methods	Training Strategy	Training Dataset	Project Code
Ren et al.	Pretrained T2V diffusion backbone + LoRA modules on temporal attention layers + appearance absorbers	One-shot motion customization from single video + appearance absorption before motion adaptation + LoRA tuning for temporal attention	Parameter-efficient LoRA fine-tuning + two-stage training (appearance absorber training → motion LoRA tuning)	Few, short videos	https://github.com/customize-a-video/customize-a-video
Bahmani et al.	Transformer-based diffusion backbone (VDiT/VD3D) + Plücker coordinate-based camera pose encoding + lightweight DiT-XS blocks	Motion spectral analysis + layer-specific camera knowledge probing + truncated normal noise schedule + feedback connections	Training with camera conditioning only in early transformer layers + standard diffusion denoising loss + truncated normal noise	Curated 20,000 video-text pairs from RealEstate10K	https://github.com/snap-research/ac3d
Ma et al.	Zero-initialized convolutional pose encoder + Pretrained text-to-image diffusion backbone + temporal & cross-frame self-attention blocks	Learnable temporal attention for motion coherence + preservation of pretrained T2I's editing ability	Two-stage training: training on image-pose pairs → finetuning on pose-free videos + minimal tuning of pretrained backbone	Curated 20,000 video-text pairs from RealEstate10K	https://github.com/mayuelala/FollowYourPose
Guo et al.	Condition encoder (shared backbone + modality heads) + frozen diffusion T2V model (AnimateDiff)	Sparse temporal control with condition propagation + masking-based sparsity simulation + purging noised ControlNet inputs + multimodal control support	Training of encoder only + freezing T2V backbone	WebVid-10M	https://github.com/guoyww/AnimateDiff
Wu et al.	Stable Video Diffusion backbone (3D U-Net) + entity representation + conditional denoising autoencoder	Segmentation tool (SAM) + 2D Gaussian creation + user trajectory input + Co-Tracker for trajectories	Supervised training with MSE loss focused on entity regions	VIPSeg	https://github.com/showlab/DragAnything

Continued on next page

 doi:10.20944/preprints202512.1476.v1

Reference	Model Architecture	Methods	Training Strategy	Training Dataset	Project Code
Xing et al.	Pretrained T2V diffusion backbone + CLIP image encoder + query Transformer + gated fusion mechanism with image/text conditioning	Dual-stream image injection (text-aligned context + visual detail guidance) + generative frame interpolation + looping videos	Three-stage training: training the image context network → adapting with T2V → joint fine-tuning with VDG	WebVid-10M	https://github.com/Doubiiu/DynamiCrafter
Jiang et al.	Pretrained T2V diffusion backbone + CLIP image encoder + attention injection module + cross-frame and temporal attention layers	Hierarchical image prompt embedding + attention injection into cross-frame attention layers + conditioning on text and image jointly	Two-stage coarse-to-fine training: training MLP encoder → training attention injection module	WebVid-10M	https://github.com/Vchitect/VideoBooth
Cai et al.	Multi-Modal Diffusion Transformer (MM-DiT) backbone with 3D full attention	Mask-guided attention sharing + latent blending + prompt token reweighting	No training or fine-tuning	None	https://github.com/TencentARC/DiTCtrl
Wei et al.	Pretrained video diffusion backbone U-Net + image retention branch + convolutional image feature extractor	Image-to-video generation + low-level image feature concatenation + double-condition guidance	Two-stage training: training of the identity adapter → fine-tuning of the motion adapter	Pexels 300K	https://github.com/ali-vilab/Vgen
Yuan et al.	Diffusion Transformer (DiT) backbone + global facial extractor + local facial extractor	Frequency-aware identity control with LF/HF features + dynamic mask loss (face) + dynamic cross-face loss	Hierarchical frequency-aware training + joint optimization of facial extractors with DiT backbone	Large-scale human face video datasets (unnamed)	https://github.com/PKU-YuanGroup/ConsisID

 doi:10.20944/preprints202512.1476.v1

4.1.3. Video Style Transfer & Editing

Video style transfer & editing involve applying the visual style or semantic modifications from a reference video to a target video, while preserving temporal coherence and motion dynamics. Unlike image-based style transfer, videos must ensure consistent frame-to-frame continuity, avoiding flicker or visual artifacts. For example, Zhang et al. [80] proposed **CAMEL**, a text-driven video editing framework built on a pretrained latent video diffusion model, which improves motion coherence and visual consistency by disentangling motion dynamics from appearance content. A key innovation is the introduction of motion prompts—optimized embeddings that capture motion characteristics from template videos. These are integrated into the latent space of diffusion models, guiding the editing process to preserve motion fidelity even under textual alterations. To further support temporal consistency, CAMEL incorporates a causal motion-enhanced attention mechanism (CAM-Attn) and a causal motion filter to isolate high-frequency motion features from low-frequency appearance details. Drawing from wavelet transform principles, it decomposes video sequences into distinct frequency bands in latent space, enabling fine-grained control over motion and appearance components. This disentangled representation allows for localized motion refinement and better appearance generalization, addressing limitations of earlier models that tightly couple motion and appearance.

By contrast, Zhang et al. [81] introduced **Style-A-Video**, a zero-shot video stylization framework that enables arbitrary text-guided style transfer while preserving the content structure and ensuring temporal coherence. Unlike previous methods that rely on style-specific training, Style-A-Video requires no additional fine-tuning, allowing flexible adaptation to diverse textual style prompts. The framework combines a generative pretrained transformer with an image latent diffusion model to produce high-fidelity, temporally coherent stylized videos directly from text prompts. To balance stylization and structural fidelity, the model refines guidance conditioning during the denoising process, preserving essential content features while applying the desired artistic transformations. Additionally, a temporal consistency module, coupled with optimized sampling strategies, further reduces inter-frame flicker and improves visual continuity.

4.1.4. Video Quality & Inference Enhancement

Complementing generation-focused approaches, some efforts have emerged to address post-generation refinement by improving the perceptual clarity, fidelity, and temporal coherence of videos. Choi et al. [23] introduced **NeuS-E**, a neuro-symbolic enhancement pipeline designed to improve the semantic and temporal coherence of videos produced by existing T2V models, without requiring additional training or architectural changes. Unlike conventional approaches that enhance generation quality through model redesign or retraining, NeuS-E operates independently as a neuro-symbolic feedback system. The method first converts text prompts into formal Temporal Logic (TL) specifications, capturing complex temporal conditions such as event sequences or motion patterns. The video itself is abstracted into an automaton representation using a Vision-Language Model (VLM), where frames are modeled as states labeled with content-derived predicates (e.g., abrupt object appearances, implausible interactions, or contradictory scene transitions), enabling formal reasoning about the video's structure. Using probabilistic model checking tools such as STORM, NeuS-V verifies whether the video automaton satisfies the TL specification, outputting a confidence score that quantifies the degree of semantic and temporal alignment between the video and the input text prompt. Finally, the detected semantic inconsistencies (e.g., abrupt object appearances, implausible interactions, or contradictory scene transitions) are fed back to guide targeted video edits, selectively modifying problematic frames or sequences to improve temporal consistency and text-video alignment.

Along similar lines of improving generation quality without retraining, Si et al. [56] introduced **FreeU**, a lightweight method that improves the quality of diffusion-based generative models by rebalancing the roles of the U-Net's backbone and skip connections. The authors observe that the backbone is primarily responsible for semantic denoising and high-level content generation, whereas the skip connections mainly inject high-frequency details during decoding but can overwhelm semantic struc-

 doi:10.20944/preprints202512.1476.v1

ture. FreeU introduces a training-free inference-time adjustment that re-weights the influence of the backbone and skip connections using just two scaling factors. This lightweight modification improves generation fidelity without altering model parameters, retraining, or increasing computational cost.

In response to inference speed limitations, Liu et al. [43] introduced **TeaCache**, a training-free, model-agnostic caching method designed to accelerate inference in video diffusion models without sacrificing visual quality. While video diffusion models produce high-fidelity outputs, their sequential denoising process leads to slow inference. Prior caching strategies, which store intermediate outputs at uniformly spaced timesteps, overlook the non-uniformity of model output changes over time—limiting both efficiency and quality preservation. TeaCache overcomes this by shifting focus from caching outputs to estimating output variation using timestep embeddings, which modulate noisy inputs in the diffusion process. By analyzing these modulated inputs, TeaCache can efficiently approximate how much the model output would change between timesteps, guiding selective caching and reuse only when differences are minimal—thus avoiding redundant computation. To enhance estimation accuracy, TeaCache incorporates a rescaling strategy that refines variation estimates without additional model calls.

Addressing physical realism in generated videos, Li et al. [38] presented **PhyT2V**, a model-agnostic, inference-time framework that improves the physical realism of videos generated by any T2V diffusion model—without retraining or additional data. Instead of introducing a new backbone, PhyT2V operates as a plug-and-play module compatible with models like Stable Video Diffusion and ModelScope T2V. It leverages off-the-shelf LLMs (e.g., GPT-4) and video captioners (e.g., Tarsier) to guide a self-refinement loop that iteratively enforces physical plausibility in generated outputs. The process begins with the LLM parsing the original user prompt to extract entities and implicit physical rules. A candidate video is then generated and automatically captioned to summarize its visual content. The LLM compares expected physical behaviors (e.g., gravity, object interactions, motion continuity) with the observed content, identifies violations (e.g., floating objects when gravity should apply), and refines the prompt to include clearer physical constraints. This loop—including rule extraction, mismatch detection, and prompt refinement—repeats until the output achieves satisfactory physical fidelity. Using chain-of-thought reasoning, the LLM can diagnose issues and suggest corrections based on commonsense physics, even if the original prompt lacks explicit cues.

In parallel, Yuan et al. [78] presented **Inflation With Diffusion**, a diffusion-based framework for T2V super-resolution that adapts pretrained image diffusion models to generate temporally coherent videos. The method "inflates" a pretrained text-to-image super-resolution diffusion model (e.g., Imagen 8× SR) into a video generation framework. This is achieved by reusing architectural components (e.g., residual blocks, cross-attention layers) to jointly process multiple video frames, thereby transferring spatial priors learned from image data without requiring extensive video-specific training. The architecture is built on a video UNet, with residual and cross-attention modules sharing weights across frames to ensure parameter efficiency and preserve spatial fidelity. A lightweight temporal adapter is introduced to model inter-frame dynamics, enabling temporal consistency without sacrificing the spatial quality inherited from the original image model.

Finally, Yang et al. [73] introduced **IPO**, a post-training framework that improves T2V generation by iteratively incorporating human preferences. In contrast to traditional methods based on one-shot or large-scale supervised fine-tuning, IPO refines T2V models through multiple optimization rounds using preference feedback—both pairwise comparisons and pointwise quality scores—to better align generated outputs with user expectations. It constructs a preference dataset with pairwise rankings (comparing two videos) and binary quality labels, enriched via LLM-augmented prompts covering categories like humans, animals, and actions. To automate the labeling process, IPO trains a critic model based on MLLMs, instruction-tuned to evaluate videos without requiring additional manual annotation or retraining. In each optimization cycle, the T2V model generates videos for given prompts, and the critic evaluates them based on human-aligned metrics such as semantic relevance, motion smoothness, subject consistency, and aesthetic quality. The model is then fine-tuned using either Direct

 doi:10.20944/preprints202512.1476.v1

Preference Optimization (DPO) for pairwise rankings or Kahneman-Tversky Optimization (KTO) for pointwise scores. This iterative loop gradually improves alignment with inferred user preferences.

4.2. Video Understanding

Video understanding involves interpreting and analyzing the content of a video to support applications such as object segmentation, activity recognition, tracking, and high-level video reasoning. For example, object segmentation consists of defining and labeling objects within each frame to enable accurate localization and shape representation. In the T2V context, such capabilities are essential for guiding generation quality, as they allow models to verify that generated scenes contain the correct objects, maintain spatial accuracy, and preserve object boundaries across frames.

To this end, Cuttano et al. [24] introduced **SAMWISE**, a lightweight extension to the Segment Anything 2 (SAM2) framework [90], developed to enable streaming-capable referring video object segmentation (RVOS) by integrating natural language understanding and explicit temporal modeling. Built on the frozen SAM2 backbone—which includes a hierarchical Vision Transformer (Hiera) visual encoder, a prompt encoder, and a mask decoder with a memory bank for temporal continuity—SAMWISE augments both visual and textual encoders with a novel *Cross-Modal Temporal* (CMT) adapter. To mitigate SAM2's tracking bias, whereby attention remains fixed on previously tracked objects even when new objects become relevant based on updated textual prompts, SAMWISE incorporates a learnable tracking bias adjustment mechanism that dynamically recalibrates attention across frames, enabling more accurate and flexible alignment with prompt-referred targets. Furthermore, it incorporates an adapter module that injects multimodal and temporal cues into the SAM2 feature extraction pipeline, without modifying the pretrained backbone or offloading cross-modal reasoning to external vision-language models.

While SAMWISE builds on a vision-first segmentation backbone, Zhu et al. [86] explored an alternative direction: leveraging pretrained T2V diffusion models themselves as frozen backbones for video understanding. Focusing on referring video object segmentation (R-VOS), they hypothesize—and empirically confirmed—that these generative models encode rich, temporally coherent, and semantically aligned latent representations useful for language-conditioned segmentation. Their framework, **VD-IT**, repurposes the latent outputs of the open-source ModelScopeT2V—a transformer-based U-Net trained for T2V synthesis—as video feature embeddings, avoiding costly full-model fine-tuning. VD-IT introduces two modules to adapt the frozen T2V backbone for R-VOS. First, the Text-Guided Image Projection module fuses text embeddings (from the T2V's native text encoder) with CLIP-extracted visual tokens to produce semantically enriched frame-level representations. Second, the Video-Specific Noise Prediction Module learns video-adaptive residuals to improve reverse diffusion and feature quality. These embeddings, along with the fused multimodal features, are passed to a mask decoder that leverages the temporal coherence of the latent space to produce spatially precise, frame-wise binary segmentation masks. Only the newly introduced components—the projection module, noise predictor, and mask decoder—are trained, using binary cross-entropy and a temporal smoothness regularization.

Extending beyond segmentation, Chen et al. [21] introduced **ShareGPT4Video**, a large-scale initiative aimed at advancing both video understanding for large video-language models (LVLMs) and T2V by providing dense, high-quality video captions. The project consists of three key components: (1) ShareGPT4Video Dataset: A collection of 40,000 videos annotated with high-precision captions generated using GPT-4V. These captions cover complex temporal dynamics, object properties, spatial relationships, camera motions, and detailed event descriptions. A novel differential captioning strategy enables stable and scalable annotation across videos of varying lengths and resolutions, addressing fine-grained intra-frame details and inter-frame temporal changes. (2) ShareCaptioner-Video: A scalable video captioning system trained on the ShareGPT4Video dataset. It has produced over 4.8 million diverse, aesthetic, and semantically rich captions, significantly expanding the pool of high-quality video–text pairs for training LVLMs and T2V models. (3) ShareGPT4Video-8B: A large multimodal

 doi:10.20944/preprints202512.1476.v1

video-language model trained on the above resources, achieving state-of-the-art results across multiple video understanding benchmarks.

4.3. Objective Assessment & Benchmark

Alongside the preceding sections on T2V generation and video understanding, parallel research efforts have focused to create curated datasets and standardized evaluation protocols to benchmark T2V models. This category is divided into two subcategories—datasets and evaluation protocols—summarized in Figure 9 and discussed in detail in the following sections.

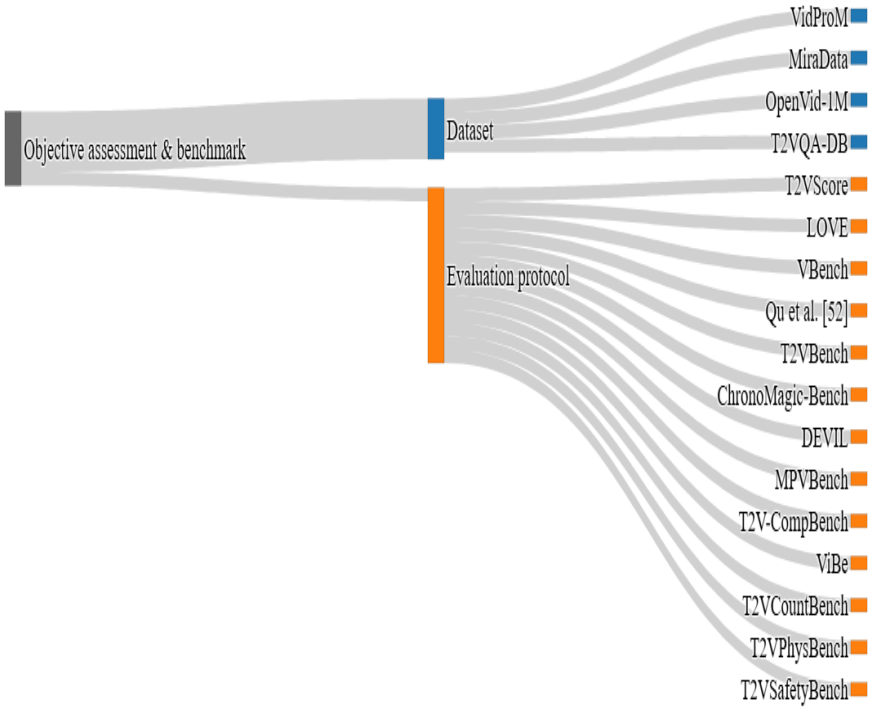

Figure 9. Schematic overview of works classified within the objective assessment & benchmark category.

4.3.1. Dataset

Researchers increasingly recognize that progress in T2V hinges on open datasets combining varied, real-world video content with detailed textual descriptions. For instance, Wang and Yang [61] introduced **VidProM**, the first large-scale dataset specifically designed to capture real user prompts and corresponding outputs for T2V diffusion models. The dataset comprises 1.67 million unique prompts sourced from real users, reflecting authentic input styles and preferences. Paired with these prompts are 6.69 million generated videos produced by four state-of-the-art diffusion models—Pika, Text2Video-Zero, VideoCraft2, and ModelScope—allowing for systematic cross-model evaluation. Each instance includes detailed metadata: the original prompt, a UUID for tracking, a timestamp, semantic embeddings (3072-dimensional vectors from OpenAI's text-embedding-3-large model), and toxicity scores across six safety categories generated via Detoxify, supporting both benchmarking and research into ethical, safe T2V generation.

In parallel, Ju et al. [36] introduced **MiraData**, a high-quality, large-scale video dataset specifically curated to address key limitations in existing video datasets, such as short durations, weak motion, and limited caption detail. MiraData features long-form clips averaging 72 seconds and richly structured captions that provide a hierarchical, multi-perspective understanding of each video. The dataset is carefully collected from diverse, manually selected sources including YouTube and other publicly

 doi:10.20944/preprints202512.1476.v1

available collections (e.g., HD-VILA-100M, Videovo, Pixabay). Each video is paired with six caption types from general overviews to fine-grained descriptions. A Short Caption summarizing the video (generated by Panda-70M), a Dense Caption capturing detailed temporal and semantic content, and Main Object/Background Captions describing primary subjects, their actions, surroundings, and environments. Additional captions include a Camera Movement Caption explaining dynamic shifts in framing, and a Style Caption detailing visual aesthetics and cinematic elements. These captions, averaging over 200 words, far exceed the descriptive capacity of prior datasets. To enable thorough evaluation, the authors introduce **MiraBench**. This benchmark suite includes 150 prompts and 17 metrics designed to assess temporal consistency, motion strength, 3D coherence, visual quality, text-video alignment, and distribution similarity in generated videos.

Lastly, Nan et al. [49] introduced **OpenVid-1M**, a large-scale, high-quality dataset designed to address major limitations in T2V generation. Unlike commonly used datasets such as WebVid-10M and Panda-70M, that suffer from low visual fidelity or high computational demands, OpenVid-1M provides over 1 million diverse, high-resolution videos paired with rich, descriptive captions. The dataset is curated based on an automated framework that emphasizes aesthetics, temporal consistency, and clarity, ensuring that each video depicts a coherent scene. To support research on high-definition video generation, OpenVid-1M also includes a selected subset called **OpenVidHD-0.4M**, which contains 433,000 videos at 1080p resolution. To demonstrate its utility, the authors develop the Multi-modal Video Diffusion Transformer (MVDiT), which jointly leverages structural features from visual tokens and semantic information from text tokens within a video diffusion backbone. Experimental results demonstrate that training on OpenVid-1M leads to superior performance over existing methods.

4.3.2. Evaluation Protocol

Early evaluation efforts generally focus on individual aspects such as temporal consistency or content continuity, using simple metrics like Fréchet Inception Distance (FID), Inception Score (IS), or CLIP similarity. However, these metrics often fail to capture the full complexity of generated videos, particularly dynamic temporal behavior and text-video alignment. Consequently, many works have moved toward human-aligned and task-specific evaluation protocols that provide a more diagnostic framework for T2V research.

For example, Wu et al. [68] proposed **T2VScore**, an evaluation framework accompanied by the **TVGE** dataset, a human-annotated benchmark. T2VScore provides a metric that jointly evaluates text-video alignment and video quality, including spatial resolution, temporal coherence, and overall visual appeal. It employs a mixture-of-experts design, combining multiple sub-metrics and specialized evaluators to to better mimic human judgment in these dimensions. To support benchmarking and validate the effectiveness of T2VScore, the authors introduce the TVGE dataset, comprising 2,543 T2V samples annotated by human evaluators along both alignment and quality axes. Experimental results on TVGE show that T2VScore correlates more strongly with human evaluations than traditional metrics, demonstrating its superiority in capturing both semantic fidelity and video quality.

Scaling this human-grounded evaluation, Wang et al. [63] introduced **LOVE**, an LMM-based evaluation framework built around AIGVE-60K—the largest and most detailed human-annotated corpus for video evaluation to date. AIGVE-60K comprises 58,500 AI-generated videos from 30 state-of-the-art T2V models, each responding to one of 3,050 diverse prompts spanning 20 fine-grained task dimensions. It includes 2.6 million human-provided annotations, such as 120,000 mean opinion scores (MOSs) evaluating perceptual video quality and 60,000 QA pairs measuring semantic-text alignment and task-specific accuracy. The LOVE framework leverages dual video encoders to extract rich spatiotemporal representations and incorporates an LLM backbone, enhanced via instruction tuning and LoRA adaptation.

Similarly, Kou et al. [37] addressed the critical need for human-aligned evaluation of T2V generative models by introducing the largest and most rigorously annotated T2V evaluation dataset, along with a novel quality assessment metric. Their contributions include the **T2V Quality Assessment DataBase (T2VQA-DB)**, comprising 10,000 videos generated by nine state-of-the-art T2V models

 doi:10.20944/preprints202512.1476.v1

using 1,000 diverse prompts selected through graph-based clustering from over a million candidates. Each video is standardized (512×512 resolution, 16 frames, 4 fps) and rated by 27 human annotators, resulting reliable mean opinion scores (MOS) that follow international annotation standards. They also introduce the **T2V Quality Assessment (T2VQA) model**, a transformer-based metric designed to predict perceptual video quality in alignment with human judgment. It jointly analyzes text-video alignment—assessing semantic relevance to the prompt—and video fidelity—capturing visual sharpness, coherence, and artifact presence.

Huang et al. [33] complemented these human-centred efforts with **VBench**, a hierarchical benchmark suite designed to evaluate the quality of modern video generative models across a diverse set of fine-grained, disentangled dimensions. Addressing the limitations of existing metrics—which often misalign with human perception and provide limited diagnostic value—VBench decomposes "video generation quality" into 16 interpretable dimensions, including subject identity consistency, motion smoothness, temporal flickering, and spatial relationships. VBench incorporates human preference annotations to better align with perceptual quality.

In parallel, Qu et al. [52] presented a framework for evaluating the quality of generated videos, addressing limitations of traditional metrics in capturing inconsistencies, semantic drift, and stylistic variation. Their approach is structured around three key dimensions: visual harmony, video-text consistency, and domain distribution gap. Visual harmony is evaluated using DOVER, a no-reference quality assessment model with aesthetic and technical branches based on ConvNeXt and Swin Transformer, respectively. These branches extract complementary spatial and temporal features that are aggregated using a learnable attention pooling to enhance perceptual sensitivity. Video-text consistency is addressed using a dual-stream architecture, where textual features guide visual alignment through a cross-attention mechanism (Text2Video Cross Attention Pooling), enabling the model to evaluate how well the video reflects the intended semantics. To capture domain distribution gaps, the model includes an auxiliary classification task that predicts the generative model source from video features, improving feature discriminability and robustness across styles and quality variations among T2V models.

While the above efforts focus on generated video quality and semantic alignment, other benchmarks prioritize temporal reasoning and motion dynamics. For instance, Wang et al. [64] introduced **T2VBench**, the first benchmark designed to evaluate the temporal reasoning capabilities of T2V generative models. While many models generate photorealistic frames, their ability to model temporal coherence, including realistic object motion, event sequencing, and camera transitions, remains underexplored. T2VBench addresses this gap with a multi-level, richly annotated evaluation suite built on 1,600+ temporally grounded prompts derived from sources such as Wikipedia. It probes a wide range of temporal phenomena, including camera movements (e.g., panning, zooming), object motions (e.g., walking, jumping), temporal event sequencing, and complex temporal interactions. The benchmark comprises more than 5,000 videos generated by state-of-the-art T2V models, each evaluated by human annotators along 16 fine-grained dimensions, capturing both low- and high-level temporal fidelity. The hierarchical evaluation enables detailed comparisons across several models; for instance, ZeroScope shows stronger performance in event dynamics, such as accurate timing and entity aggregation. However, all models still struggle with maintaining consistent temporal coherence and realistic motion.

Focusing on temporal metamorphosis, Yuan et al. [76] introduced **ChronoMagic-Bench**, a benchmark for evaluating temporal reasoning and metamorphic capabilities in T2V models, specifically within time-lapse video synthesis. Unlike prior benchmarks that emphasize visual fidelity or text-video alignment, ChronoMagic-Bench focuses on a model's ability to capture metamorphic amplitude—the extent of meaningful transformations over time—and temporal coherence, the logical progression of visual content. The benchmark includes 1,649 curated prompts, each paired with real-world reference videos, covering four major categories (biological phenomena, human-created processes, meteorological events, and physical phenomena) and 75 subcategories. To support quantitative analysis,

 doi:10.20944/preprints202512.1476.v1

the authors propose two novel automatic metrics: MTScore (Metamorphic Amplitude Score), which quantifies the magnitude and semantic meaningfulness of changes over time, and CHScore (Coherence Score), which evaluates the smoothness and plausibility of temporal transitions. The authors also introduce **ChronoMagic-Pro**, a 460K video-caption dataset (720p), tailored to physically grounded, time-lapse transformations for training and benchmarking.

Shifting focus to motion dynamics, Liao et al. [40] introduced **DEVIL**, an evaluation protocol specifically designed to evaluate video dynamics in T2V generation. The authors argue that dynamics, reflecting the intensity and fidelity of motion relative to textual prompts, are critical for evaluating how well generated videos capture the vividness implied by the input. DEVIL includes a benchmark with text prompts covering varying levels of dynamic content—from static scenes to highly dynamic actions—paired with metrics that analyze motion at different temporal granularities. This enables fine-grained evaluation of how motion unfolds over time. Three key metrics form the core of DEVIL's protocol: *Dynamics Range*, which measures the diversity and intensity of motion; *Dynamics Controllability*, which evaluates how well models adjust motion based on the prompt's intended dynamics; and *Dynamics-based Quality*, which evaluates the perceptual realism and semantic alignment of the motion.

In multi-prompt video generation, Cai et al. [20] introduced **MPVBench**, a benchmark designed to evaluate models guided by sequential or multiple textual prompts. Such models must maintain smooth, coherent transitions between prompt segments while preserving consistent object motion and overall temporal fidelity. MPVBench features 130 long-form prompts with 10 diverse transition types crafted using GPT-4. To overcome limitations of standard metrics like CLIP similarity—which often overlook abrupt semantic shifts at prompt boundaries—MPVBench proposes a novel metric called the CLIP Similarity Coefficient of Variation (CSCV). CSCV measures the uniformity of CLIP similarity scores across adjacent frames, penalizing isolated or abrupt semantic changes that indicate poor transition smoothness. Alongside CSCV, MPVBench employs text-image similarity metrics to assess prompt fidelity and incorporates motion smoothness measures from benchmarks such as VBench to evaluate physical realism and motion fluidity.

Another line of research in evaluation protocols focuses on compositional reasoning, hallucinations, counting, and physical fidelity. For instance, Sun et al. [62] introduced **T2V-CompBench**, the first benchmark specifically designed to evaluate the compositional reasoning capabilities of T2V generative models. The benchmark covers seven core compositional categories, including attribute binding, spatial relations, motion and action binding, object interactions, and generative numeracy. It features 1,400 curated prompts that probe each aspect, enabling fine-grained assessment across diverse compositional scenarios. T2V-CompBench also introduces a hybrid evaluation framework, combining multimodal large language model (MLLM)-based metrics, detection-based assessments, and motion tracking techniques. These metrics evaluate semantic consistency, spatial correctness, and motion coherence, and have been validated against human judgments for reliability.

Addressing hallucinations, Rawte et al. [53] introduced **ViBe**, the first large-scale benchmark specifically designed to evaluate and categorize hallucinations in T2V generative models. ViBe systematically defines five primary hallucination types frequently observed in T2V results: vanishing subject, numeric variability, temporal dysmorphia, omission error, and physical incongruity. These categories capture inconsistencies such as disappearing entities, incorrect object counts, unnatural temporal transitions, missing elements described in the text prompt, and physically implausible behavior. The benchmark consists of 3,782 videos generated by prompting 10 open-source T2V models with 837 diverse captions from the MS COCO dataset. Each video is manually annotated to identify hallucination types, enabling precise analysis of generative failure modes.

Focusing on numerical precision, Guo et al. [30] introduced **T2VCountBench**, a specialized human evaluation benchmark designed to evaluate the ability of ten state-of-the-art T2V generative models—including both open-source and commercial models available up to 2025—to correctly follow numerical instructions involving object counting. Despite significant advances in producing realistic, high-quality videos, the study reveals a critical limitation: current models struggle to generate videos with precise

 doi:10.20944/preprints202512.1476.v1

object counts, especially for numbers up to nine. Factors such as video style, temporal dynamics, and multilingual prompts have minimal impact on performance. Furthermore, decomposing complex counting tasks or refining prompts offers only marginal improvements, highlighting fundamental challenges in existing video generation architectures.

Addressing physical plausibility, Guo et al. [31] introduced **T2VPhysBench**, the first-principles benchmark to evaluate whether state-of-the-art T2V generative models adhere to fundamental physical laws in their outputs. T2VPhysBench systematically measures physical consistency beyond simple visual or semantic criteria, focusing on twelve core physical laws grouped into three categories: Newtonian principles (Newton's three laws and universal gravitation), conservation principles (energy, mass, linear and angular momentum), and phenomenological principles (Hooke's Law, Snell's Law, the Law of Reflection, and Bernoulli's Principle). It evaluates both open-source and commercial models released between 2023 and 2025 (e.g., OpenKling, Wan 2.1, Sora, Stable Diffusion Video) using standardized short clips (720p, 16:9, 4 seconds). Results reveal systemic shortcomings: no model consistently respects all tested physical laws, with average compliance scores below 0.60 across categories—highlighting critical gaps in temporal and physical coherence and underscoring the need for physics-aware video generation techniques.

While prior evaluations focused on video quality, concerns over unsafe, illegal, or unethical generated content have escalated. Miao et al. [47] introduced **T2VSafetyBench**, the first benchmark designed to evaluate the safety risks of T2V generative models. It provides a multifaceted safety framework containing four primary categories covering 14 critical aspects, including illegal activities, violence, hate speech, misinformation, and other harms, tailored to the spatiotemporal nature of video. The benchmark includes a diverse malicious prompt dataset drawn from real-world unsafe prompts, LLM-generated content, and jailbreak attack prompts designed to bypass filters, testing models under realistic adversarial conditions. Evaluations of nine state-of-the-art T2V models, using both human reviews and automated GPT-4 assessments, revealed a strong correlation between methods. The findings show that no single model excels across all safety dimensions, underscoring an ongoing trade-off between safety and usability.

5. Discussion

To shape and consolidate knowledge in the emerging field of T2V generation, this study synthesizes fragmented research efforts and proposed T2V models. This section will answer to the following research questions:

- RQ1: What are the key technological advances that have driven progress in the aforementioned video research fields?
- RQ2: What are the current challenges and best practices in evaluating T2V models, and how do benchmark datasets support the development of robust text-to-video generators?
- RQ3: What are the primary technical challenges and future research directions in leveraging Generative AI and LLMs for text-to-video generation?

RQ1: What are the key technological advances that have driven progress in the aforementioned video research fields?

In recent years, several foundational breakthroughs have underpinned the rapid progress in T2V generation. Among these, diffusion models have become a cornerstone since it has demonstrated unprecedented capabilities in generating high-fidelity, temporally coherent videos from text prompts. An important milestone was the release of Sora, OpenAI's diffusion-transformer model that can generate a two-minute video with remarkable realism, marking a leap forward in video synthesis. These diffusion transformers (DiTs) exemplify how transformer-based denoising architectures can maintain temporal consistency across frames, outperforming earlier approaches. At the same time, alternative generative paradigms like autoregressive Video Pixel Networks and VAE-based video generators, provided important groundwork, however, their impact has been overshadowed by the superior quality and flexibility of diffusion models. Modern T2V systems often employ hybrid architectures

 doi:10.20944/preprints202512.1476.v1

that leverage strengths from multiple paradigms; for example, latent video diffusion models combine a VAE compression of video frames with transformer-guided diffusion, exploiting spatial detail and temporal smoothness. This convergence on scalable generative architectures, particularly diffusion models enhanced by transformer backbones, has enabled the generation of videos that are both more realistic and of longer duration than was previously achievable.

Another key driver is the explosion of large-scale video–text data and robust pretraining regimes. Models now leverage massive datasets (e.g. WebVid-10M, LAION-5B, and other web-scale video-caption collections) to learn rich cross-modal representations. By pretraining on millions of diverse video–text pairs and then fine-tuning for text-to-video tasks, generative models acquire a broad "world knowledge" of visuals and language. This transfer learning paradigm is bolstered by open, high-quality datasets specifically curated for video generation research. For example, the **OpenVid-1M** dataset introduced 1 million high-resolution video clips with descriptive captions to endorse T2V training, and the **VidProM** dataset collected 1.67 million real user prompts with corresponding model-generated videos to reflect authentic use cases. The availability of such data has been pivotal as researchers increasingly recognize that progress in T2V "hinges on open datasets combining varied, real-world video content with detailed textual descriptions". In parallel, some T2V models have leveraged multimodal learning by integrating visual encoders with large language models (LLMs) to improve the semantic understanding of video content. The pretrained LLMs (e.g., GPT or LLaMA) can be employed to guide video generation or interpret prompts with nuanced context, which improves text-video alignment. For example, vision-language transformers that use both visual features and LLM-driven text embeddings enable to maintain the narrative coherence and object relevance in videos. Overall, the synergy of generative architectures (especially diffusion transformers), training on large-scale video–text corpora, and incorporation of LLM-based semantic reasoning has dramatically accelerated progress in text-to-video generation, laying a strong foundation for future state-of-the-art models.

RQ2: What are the current challenges and best practices in evaluating T2V models, and how do benchmark datasets support the development of robust text-to-video generators?

Despite rapid technical progress, evaluating T2V models remains a complex challenge. Traditional evaluation metrics inherited from image generation (e.g. FID for visual quality or CLIP-based similarity for relevance) often fail to capture the full temporal and semantic nuances of generated videos. A persistent issue is that many models can generate photorealistic individual frames but still struggle with consistency across time, yet simple metrics may not penalize subtle temporal glitches or logical breaks in a video. Early evaluation efforts tended to examine multiple aspects of quality, such as per-frame fidelity or text alignment. As a result, best practices in T2V evaluation have shifted toward multi-dimensional and human-centered approaches. For example, the benchmark **VBench** decomposes video quality into 16 fine-grained dimensions (e.g., motion smoothness, temporal consistency, object permanence, absence of flicker, spatial relationships), providing a more diagnostic assessment of generative performance beyond a single aggregate score. In general, combining automated metrics with human judgment is now considered essential: benchmarks often report human evaluation of text–video alignment and overall realism to validate whether numeric scores (e.g., from CLIP or FID) truly reflect perceptual quality. For instance, the **T2V-CompBench** suite integrates *"multimodal LLM-based metrics, detection-based assessments, and motion tracking techniques"* and validates them against human judgments to ensure reliability. Aligning metric outcomes with human perception is crucial, as purely algorithmic scores can be misleading in capturing narrative coherence or subtle errors.

Another best practice is the development of specialized evaluation benchmarks targeting known failure modes of T2V models. Researchers have started targeting capabilities like temporal reasoning, compositional understanding, and realism under physical constraints. For example, **T2VBench** focuses on temporal dynamics, providing 1,600+ annotated videos to test if models maintain coherent object motion, event sequencing, and scene transitions over time. This addresses the gap that while many models produce sharp frames, they often fail on long-term coherence and plausible scene progressions.

 doi:10.20944/preprints202512.1476.v1

Likewise, **T2V-CompBench** was introduced as the first benchmark for compositional reasoning in generated videos, covering attributes binding, spatial relations, object interactions, and even basic counting (numeracy) in complex prompts, enabling fine-grained evaluation of whether a model can correctly combine multiple concepts (e.g., "red cube on a blue sphere while a cat jumps") in one video. Results on this benchmark have revealed that current models often struggle with attributing the right properties to the right entities or maintaining multiple object relationships over time, reflecting weaker compositional capabilities compared to image generation. In response, the benchmark pairs automated checks (object detectors, spatial consistency metrics) with LLM-based semantic evaluations, striving for a holistic measure of compositional faithfulness.

Moreover, emerging benchmarks tackle even more specific challenges: **hallucination** in videos are addressed by the **ViBe** dataset, which defines five types of video hallucinations and evaluates how frequently models introduce objects or actions that were never mentioned or plausible. Early analyses show that even state-of-the-art models can noticeably hallucinate, for instance by adding extraneous elements or morphing object identities, a shortcoming that **ViBe** helps quantify. **Counting ability** is probed by a dedicated evaluation where humans judge whether models accurately represent numeric quantities in the video (e.g., "five apples"). Guo et al. [30] found that across ten contemporary T2V models, counting beyond very small numbers remains unreliable, prompting the creation of a **T2VCountBench** with human evaluation to drive improvements. **Physical realism** is another critical evaluation dimension: a recent benchmark tested various generative models against fundamental physical laws (e.g., gravity, object permanence, reflection) and found *"no model consistently respects all tested physical laws"*, with average compliance scores under 60%. This underscores that physics awareness in video generation is nascent, motivating new metrics for physical consistency and efforts to embed physical knowledge into models. Finally, **content safety** has become an evaluation focus amid growing concerns that models may generate harmful or inappropriate videos. Miao et al. [47] introduced **T2VSafetyBench**, which evaluates T2V models on 14 safety aspects (violence, hate, illicit behavior, etc.) using a suite of adversarial prompts and human/GPT-4 assessments. Initial results show no single model excels in all safety dimensions, and that there are trade-offs between minimizing unsafe outputs and retaining creative freedom. The proliferation of these benchmarks and protocols illustrates the community's best practice: use a wide spectrum of evaluation datasets to stress-test models across multiple dimensions of quality and reliability. By benchmarking on diverse tasks, from temporal coherence to ethical safety, researchers can obtain a more robust picture of a model's strengths and weaknesses and drive the development of more *"robust text-to-video generators"*, as called for in RQ2.

RQ3: What are the primary technical challenges and future research directions in leveraging Generative AI and LLMs for text-to-video generation?

Despite the notable progress in the text-to-video field, significant technical challenges still need further research. One major challenge is **scaling to longer-duration videos** without compromising temporal coherence or incurring prohibitive computation costs. Generating videos beyond a few seconds poses challenges in maintaining consistent narratives and preventing drift in visual details or motion. Current T2V models typically produce only short clips (often 2-5 seconds) at modest resolutions. Moving to minute-long or even narrative-length videos will require new strategies to handle long-range temporal dependencies. Promising directions include architectural innovations like causal or autoregressive video generation that generate frames sequentially rather than with full sequence attention (e.g, **CausVid**). This avoids needing to attend to an entire video at once and, combined with distillation techniques to reduce diffusion sampling steps. Such advances illustrate a path forward for longer video generation: hierarchical or chunked generation (splitting a long video into manageable segments) and efficient sampling will be key research areas. Additionally, even with better architectures, training models for long outputs is resource-intensive – pointing to a need for **optimized training paradigms**, perhaps using curriculum learning (start with short sequences, gradually increase length) or leveraging video continuation tasks where a model extends a given clip.

 doi:10.20944/preprints202512.1476.v1

A second major challenge is **aligning T2V generation with human preferences, intent, and safety requirements**. As models become more powerful, simply optimizing for pixel quality is not enough – they must also produce videos that are useful and trustworthy to users. This has led to the incorporation of human feedback loops and preference modeling into the generation process. For instance, *"addressing the critical challenge of aligning T2V models with diverse human preferences"* has spurred techniques like Direct Preference Optimization for video, exemplified by the **VideoDPO** framework that combines multiple criteria (visual clarity, temporal stability, text relevance) to rank outputs, improving both the fidelity and text alignment of videos as judged by users. Such approaches, akin to reinforcement learning from human feedback in text domains, are a promising direction to ensure generative videos meet qualitative expectations. Going forward, we anticipate more work on **multi-criteria reward models** (for example, jointly evaluating realism, relevance, creativity, and safety) and on efficient collection of preference data specific to video. Another aspect of alignment is **content safety**: as noted, no model currently excels at avoiding all forms of unsafe content. Future research must integrate safety constraints into generation, possibly by modeling forbidden content within the training objective or using real-time detectors to steer generation away from problematic scenes. Likewise, bias and fairness in T2V outputs (e.g., how different demographics are portrayed) will need examination as the technology matures, building on the initial safety benchmarks.

A third research direction involves enhancing the **world knowledge, reasoning, and physical realism** of generative video models. Today's models often generate superficially plausible videos that nevertheless break logical rules or physical laws upon close inspection (e.g. objects appearing/disappearing, unnatural motion, physics violations). This limitation stems from models learning correlations in training data but lacking an understanding of underlying principles. Future T2V generators will benefit from being "physics-aware & reasoning-grounded", meaning they have mechanisms to enforce basic consistency and continuity in the worlds they simulate. One promising direction is to explicitly inject knowledge of physics or common sense into the generative process like the **WISA** framework. Another approach is neuro-symbolic integration, where symbolic logic or external reasoning modules work alongside neural generators. In general, coupling LLMs with video generation is a tantalizing research avenue: LLMs could help maintain narrative coherence (by generating high-level storyboards or shot lists from a script), ensure semantic alignment (verifying that each scene logically follows the description), or augment evaluation (as GPT-4V is already used to assess video captions and safety). Early efforts like ShareGPT4Video have used GPT-4 with vision to generate detailed video annotations at scale, effectively leveraging an LLM to improve video datasets. Future systems might extend this idea, using multimodal LLMs as intelligent directors or critics in the video generation loop.

Finally, researchers are exploring **hybrid and factorized generation** strategies to overcome current limitations of end-to-end models. Rather than training a single colossal model to do everything, one direction is to split the problem into sub-tasks or stages that can be optimized independently. For example, to exploit abundantly available image data and unlabeled videos, one can decouple what to render from how it moves: a content branch generates high-quality key frames (leveraging image-text data), while a motion branch, trained on unlabeled video, ensures those frames evolve smoothly over time. Similarly, factorized diffusion models like **Show-1** break the generation into an image generation step followed by a video extension step, reusing powerful image diffusion models for efficiency. Such modular designs make it easier to scale certain components (e.g., using a very large text-to-image model for detail, and a specialized temporal model for motion). The trade-off, however, is managing consistency between components; hence research is needed on optimal interfaces between image and video generation modules (e.g., how to condition a video model on key frames or latent codes without drift). **Model scalability** is another concern: pushing parameter counts and model capacity is a clear trend (with recent T2V foundation models reaching tens of billions of parameters), but training such models is expensive. Efficient fine-tuning methods (e.g., parameter-efficient adapters or layer freezing) will be vital to make large T2V models accessible for wider use. There is also growing interest in **multi-lingual and multimodal video generation**. So far, most T2V benchmarks and models assume

 doi:10.20944/preprints202512.1476.v1

English text inputs, but recent work like Step-Video-T2V introduced dual text encoders to support bilingual generation in English and Chinese. This points toward a future where T2V systems handle many languages and cultural contexts. Along the same lines, adding other modalities such as **audio** is largely uncharted territory since current generators typically produce silent videos, a notable limitation for real-world applications. Future research should integrate text-to-speech or sound effect generation synchronized with video frames, enabling fully multimodal storytelling (e.g., a video with dialog or ambient sound matching the scene). Addressing audio generation, alongside visual content, would enable text-to-video models to generate complete immersive experiences rather than just silent movies.

6. Conclusions

This study conducted a systematic literature review of scientific publications exploring text-to-video (T2V) generation, with a specific emphasis on works published from 2024 onward that address methods, datasets, and evaluation practices for generating temporally coherent and semantically consistent video from natural language prompts. In total, 69 studies were gathered to consolidate and structure the fragmented knowledge in this emerging field. The main findings indicate that transformer architectures combined with diffusion models has lead to a remarkable progress in generating visual content from text prompts. Moreover, Large language models (LLMs) are increasingly utilized for better alignment between text prompts and videos. Another observation pertains to the diversity in evaluation practices: some studies rely only on automated metrics, while others combine human judgments, introducing variability that complicates direct comparisons across methods.

Subsequently, this review contributed to the theory by clarifying the current research landscape on T2V and proposing directions for future investigation. Specifically, the main contributions of this study are as follows: (1) mapping the main approaches in T2V generation addressed by recent research, (2) outlining the existing evaluation landscape of T2V models, and (3) providing a structured roadmap for researchers and practitioners to support further research on T2V generation.

Neverthless, it is important to acknowledge certain limitations of this research. (1) The review focuses specifically on T2V generation and does not cover related modalities such as text-to-image and image-to-video; (2) only English-language journal papers and conference proceedings were considered; and (3) although the consulted databases are regularly updated, only papers retrieved at the consultation time were reviewed. Nevertheless, by following a rigorous methodology and addressing the main research questions, this review provides a consolidated foundation for future work in T2V generation.

Author Contributions: Conceptualization, K.H. and S.S.; methodology, K.H. and S.S.; software, not applicable; validation, K.H. and S.S.; formal analysis, K.H.; investigation, K.H. and S.S.; resources, K.H. and S.S.; data curation, K.H.; writing—original draft preparation, K.H.; writing—review and editing, K.H. and S.S.; visualization, K.H.; supervision, S.S.; project administration, K.H.; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest..

 doi:10.20944/preprints202512.1476.v1

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
T2V	Text-to-video
DiT	Diffusion Transformer
VAE	Variational Autoencoders
SVG	Stochastic Video Generation
LLM	Large Language Models
LVLM	Large video-language models
VDM	Video Diffusion Models
VPN	Video Pixel Networks
LSTM	Long Short-Term Memory
LoRA	Low-Rank Adaptation
U-ViT	U-shaped Vision Transforme
AdaLN	Adaptive LayerNorm
DPO	Direct Preference Optimization
FID	Fréchet Inception Distance
IS	Inception Score

References

1. Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; et al. Video Generation Models as World Simulators. https://openai.com/index/video-generation-models-as-world-simulators/, 2024.
2. Kalchbrenner, N.; van den Oord, A.; Simonyan, K.; Danihelka, I.; Vinyals, O.; Graves, A.; Kavukcuoglu, K. Video Pixel Networks. In Proceedings of the Proceedings of the 34th International Conference on Machine Learning (ICML). PMLR, 2017, Vol. 70, *Proceedings of Machine Learning Research*, pp. 1771–1779. https://doi.org/10.5555/3305381.3305564.
3. Denton, E.; Fergus, R. Stochastic Video Generation with a Learned Prior. In Proceedings of the Proceedings of the 35th International Conference on Machine Learning (ICML). PMLR, 2018, Vol. 80, *Proceedings of Machine Learning Research*, pp. 1174–1183. https://doi.org/10.5555/3294996.3295094.
4. OpenAI. GPT-4 Technical Report **2023**.
5. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* **2023**.
6. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bikel, D.; Blecher, L.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288* **2023**.
7. Lin, B.; Zhu, B.; Ye, Y.; Ning, M.; Jin, P.; Yuan, L. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv preprint arXiv:2311.10122* **2023**.
8. OpenAI. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/, 2024. Describes GPT-4o, a multimodal large language model with text, audio, and vision capabilities.
9. Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; Fleet, D.J. Video Diffusion Models. *arXiv preprint arXiv:2204.03458* **2022**.
10. van den Oord, A.; Kalchbrenner, N.; Vinyals, O.; Espeholt, L.; Graves, A.; Kavukcuoglu, K. Conditional Image Generation with PixelCNN Decoders. *arXiv preprint arXiv:1606.05328* **2016**.
11. Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Kaczmarczyk, R.; Schaeffer, K.; Shah, S.A.; et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In Proceedings of the NeurIPS Datasets and Benchmarks Track, 2022.
12. Chen, T.S.; Siarohin, A.; Menapace, W.; Deyneka, E.; wei Chao, H.; Jeon, B.E.; Fang, Y.; Lee, H.Y.; Ren, J.; Yang, M.H.; et al. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13320–13331. CVPR 2024.

 doi:10.20944/preprints202512.1476.v1

13. Bain, M.; Zhu, A.; Sidorov, E.; Laurens, V.; Holden, D.; et al. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. *arXiv preprint arXiv:2104.00650* **2021**.
14. Bain, M.; Nagrani, A.; Varol, G.; Zisserman, A. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1728–1738. https://doi.org/10.1109/CVPR46437.2021.00181.
15. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685, 2022. Accessed: 2025-07-21.
16. Liu, Y.; Agarwal, S.; Venkataraman, S. AutoFreeze: Automatically Freezing Model Blocks to Accelerate Fine-tuning. *arXiv preprint arXiv:2102.01386* **2021**.
17. Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. Make-A-Video: Text-to-Video Generation without Text-Video Data. *arXiv preprint arXiv:2209.14792* **2022**.
18. Bahmani, S.; Skorokhodov, I.; Qian, G.; Siarohin, A.; Menapace, W.; Tagliasacchi, A.; Lindell, D.B.; Tulyakov, S. AC3D: Analyzing and Improving 3D Camera Control in Video Diffusion Transformers. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
19. Bao, F.; Xiang, C.; Yue, G.; He, G.; Zhu, H.; Zheng, K.; Zhao, M.; Liu, S.; Wang, Y.; Jun, Z. Vidu: A Highly Consistent, Dynamic and Skilled Text-to-Video Generator with Diffusion Models. *arXiv preprint arXiv:2405.04233* **2024**.
20. Cai, M.; Cun, X.; Li, X.; Liu, W.; Zhang, Z.; Zhang, Y.; Shan, Y.; Yue, X. DiTCtrl: Exploring Attention Control in Multi-Modal Diffusion Transformer for Tuning-Free Multi-Prompt Longer Video Generation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2025, pp. 7763–7772.
21. Chen, L.; Wei, X.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Lin, B.; Tang, Z.; et al. ShareGPT4Video: Improving Video Understanding and Generation with Better Captions. In Proceedings of the Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track, 2024.
22. Chen, H.; Zhang, Y.; Cun, X.; Xia, M.; Wang, X.; Weng, C.; Shan, Y. VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2024, pp. 7310–7320.
23. Choi, M.; Sharan, S.P.; Goel, H.; Shah, S.; Chinchali, S. We'll Fix it in Post: Improving Text-to-Video Generation with Neuro-Symbolic Feedback. *arXiv preprint arXiv:2504.17180* **2025**.
24. Cuttano, C.; Trivigno, G.; Rosi, G.; Masone, C.; Averta, G. SAMWISE: Infusing Wisdom in SAM2 for Text-Driven Video Segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
25. Dalal, K.; Koceja, D.; Hussein, G.; Xu, J.; Zhao, Y.; Song, Y.; Han, S.; Cheung, K.C.; Kautz, J.; Guestrin, C.; et al. One-Minute Video Generation with Test-Time Training. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
26. Fei, H.; Wu, S.; Ji, W.; Zhang, H.; Chua, T.S. Dysen-VDM: Empowering Dynamics-aware Text-to-Video Diffusion with LLMs. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2024, pp. 7641–7653.
27. Gal, R.; Vinker, Y.; Alaluf, Y.; Bermano, A.H.; Cohen-Or, D.; Shamir, A.; Chechik, G. Breathing Life Into Sketches Using Text-to-Video Priors. *arXiv preprint arXiv:2311.13608* **2023**. https://doi.org/10.48550/arXiv.2311.13608.
28. Girdhar, R.; Singh, M.; Brown, A.; Duval, Q.; Azadi, S.; Rambhatla, S.S.; Shah, A.; Yin, X.; Parikh, D.; Misra, I. Factorizing Text-to-Video Generation by Explicit Image Conditioning. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2024. https://doi.org/10.1007/978-3-031-73033-7_12.
29. Guo, Y.; Yang, C.; Rao, A.; Agrawala, M.; Lin, D.; Dai, B. SparseCtrl: Adding Sparse Controls to Text-to-Video Diffusion Models. In Proceedings of the European Conference on Computer Vision (ECCV), 2024.
30. Guo, X.; Huang, Z.; Huo, J.; Liang, Y.; Shi, Z.; Song, Z.; Zhang, J. Can You Count to Nine? A Human Evaluation Benchmark for Counting Limits in Modern Text-to-Video Models. *arXiv preprint arXiv:2504.04051* **2025**.
31. Guo, X.; Huo, J.; Shi, Z.; Song, Z.; Zhang, J.; Zhao, J. T2VPhysBench: A First-Principles Benchmark for Physical Consistency in Text-to-Video Generation. *arXiv preprint arXiv:2505.00337* **2025**.
32. Henschel, R.; Khachatryan, L.; Poghosyan, H.; Hayrapetyan, D.; Tadevosyan, V.; Wang, Z.; Navasardyan, S.; Shi, H. StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text. In

 doi:10.20944/preprints202512.1476.v1

Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025. https://doi.org/10.48550/arXiv.2403.14773.

33. Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. VBench: Comprehensive Benchmark Suite for Video Generative Models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2024.
34. Jeong, H.; Park, G.Y.; Ye, J.C. VMC: Video Motion Customization using Temporal Attention Adaption for Text-to-Video Diffusion Models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2024, pp. 9212–9221.
35. Jiang, Y.; Wu, T.; Yang, S.; Si, C.; Lin, D.; Qiao, Y.; Loy, C.C.; Liu, Z. VideoBooth: Diffusion-based Video Generation with Image Prompts. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2024, pp. 6689–6699.
36. Ju, X.; Gao, Y.; Zhang, Z.; Yuan, Z.; Wang, X.; Zeng, A.; Xiong, Y.; Xu, Q.; Shan, Y. MiraData: A Large-Scale Video Dataset with Long Durations and Structured Captions, 2024, [arXiv:cs.CV/2407.06358].
37. Kou, T.; Liu, X.; Zhang, Z.; Li, C.; Wu, H.; Min, X.; Zhai, G.; Liu, N. Subjective-Aligned Dataset and Metric for Text-to-Video Quality Assessment. In Proceedings of the Proceedings of the 2024 International Conference on Machine Learning (ICML) Workshop or similar venue, 2024.
38. Li, J.; Yu, S.; Lin, H.; Cho, J.; Yoon, J.; Bansal, M. PhyT2V: LLM-Guided Iterative Self-Refinement for Physics-Grounded Text-to-Video Generation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2024.
39. Li, J.; Yu, S.; Lin, H.; Cho, J.; Yoon, J.; Bansal, M. Training-free Guidance in Text-to-Video Generation via Multimodal Planning and Structured Noise Initialization. *arXiv preprint arXiv:2504.08641* **2025**. https://doi.org/10.48550/arXiv.2504.08641.
40. Liao, M.; Lu, H.; Zhang, X.; Wan, F.; Wang, T.; Zhao, Y.; Zuo, W.; Ye, Q.; Wang, J. Evaluation of Text-to-Video Generation Models: A Dynamics Perspective. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) 37, 2024.
41. Lin, B.; Ge, Y.; Cheng, X.; Li, Z.; Zhu, B.; Wang, S.; He, X.; Ye, Y.; Yuan, S.; Chen, L.; et al. Open-Sora Plan: Open-Source Large Video Generation Model. *arXiv preprint arXiv:2412.00131* **2024**.
42. Liu, R.; Wu, H.; Zheng, Z.; Wei, C.; He, Y.; Pi, R.; Chen, Q. VideoDPO: Omni-Preference Alignment for Video Diffusion Generation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
43. Liu, F.; Zhang, S.; Wang, X.; Wei, Y.; Qiu, H.; Zhao, Y.; Zhang, Y.; Ye, Q.; Wan, F. Timestep Embedding Tells: It's Time to Cache for Video Diffusion Model. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
44. Lv, J.; Huang, Y.; Yan, M.; Huang, J.; Liu, J.; Liu, Y.; Wen, Y.; Chen, X.; Chen, S. GPT4Motion: Scripting Physical Motions in Text-to-Video Generation via Blender-Oriented GPT Planning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2024. https://doi.org/10.48550/arXiv.2311.12631.
45. Ma, Y.; He, Y.; Cun, X.; Wang, X.; Chen, S.; Shan, Y.; Li, X.; Chen, Q. Follow Your Pose: Pose-Guided Text-to-Video Generation Using Pose-Free Videos. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2024, Vol. 38, pp. 4117–4125. https://doi.org/10.1609/aaai.v38i5.28206.
46. Menapace, W.; Siarohin, A.; Skorokhodov, I.; Deyneka, E.; Chen, T.S.; Kag, A.; Fang, Y.; Stoliar, A.; Ricci, E.; Ren, J.; et al. Snap Video: Scaled Spatiotemporal Transformers for Text-to-Video Synthesis. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16820–16830. https://doi.org/10.1109/CVPR56347.2024.01639.
47. Miao, Y.; Zhu, Y.; Dong, Y.; Yu, L.; Zhu, J.; Gao, X.S. T2VSafetyBench: Evaluating the Safety of Text-to-Video Generative Models. In Proceedings of the Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track, 2024.
48. Mohamed, A.A.; Lucke-Wold, B. Text-to-video generative artificial intelligence: Sora in neurosurgery. *Neurosurgical Review* **2024**, *47*, 272. https://doi.org/10.1007/s10143-024-02514-w.
49. Nan, K.; Xie, R.; Zhou, P.; Fan, T.; Yang, Z.; Chen, Z.; Li, X.; Yang, J.; Tai, Y. OpenVid-1M: A Large-Scale High-Quality Dataset for Text-to-Video Generation. In Proceedings of the International Conference on Learning Representations (ICLR), 2025.
50. Qin, C.; Xia, C.; Ramakrishnan, K.; Ryoo, M.S.; Tu, L.; Feng, Y.; Shu, M.; Zhou, H.; Awadalla, A.; Wang, J.; et al. xGen-VideoSyn-1: High-Fidelity Text-to-Video Synthesis with Compressed Representations. *arXiv preprint arXiv:2408.12590* **2024**.

 doi:10.20944/preprints202512.1476.v1

51. Qing, Z.; Zhang, S.; Wang, J.; Wang, X.; Wei, Y.; Zhang, Y.; Gao, C.; Sang, N. Hierarchical Spatio-temporal Decoupling for Text-to-Video Generation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6334–6344.
52. Qu, B.; Liang, X.; Sun, S.; Gao, W. Exploring AIGC Video Quality: A Focus on Visual Harmony, Video-Text Consistency and Domain Distribution Gap. *arXiv preprint arXiv:2404.13573* **2024**.
53. Rawte, V.; Jain, S.; Sinha, A.; Kaushik, G.; Bansal, A.; Vishwanath, P.R.; Jain, S.R.; Reganti, A.N.; Jain, V.; Chadha, A.; et al. ViBe: A Text-to-Video Benchmark for Evaluating Hallucination in Large Multimodal Models. *arXiv preprint arXiv:2411.10867* **2024**.
54. Ren, Y.; Zhou, Y.; Yang, J.; Shi, J.; Liu, D.; Liu, F.; Kwon, M.; Shrivastava, A. Customize-A-Video: One-Shot Motion Customization of Text-to-Video Diffusion Models. In Proceedings of the European Conference on Computer Vision (ECCV), 2024.
55. Sharan, S.P.; Choi, M.; Shah, S.; Goel, H.; Omama, M.; Chinchali, S. Neuro-Symbolic Evaluation of Text-to-Video Models using Formal Verification. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
56. Si, C.; Huang, Z.; Jiang, Y.; Liu, Z. FreeU: Free Lunch in Diffusion U-Net. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 184–193.
57. Tan, S.; Gong, B.; Feng, Y.; Zheng, K.; Zheng, D.; Shi, S.; Shen, Y.; Chen, J.; Yang, M. Mimir: Improving Video Diffusion Models for Precise Text Understanding. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2025, pp. 23978–23988.
58. Tian, Y.; Yang, L.; Yang, H.; Gao, Y.; Deng, Y.; Chen, J.; Wang, X.; Yu, Z.; Tao, X.; Wan, P.; et al. VideoTetris: Towards Compositional Text-to-Video Generation. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) 37, 2024.
59. Wang, X.; Zhang, S.; Yuan, H.; Qing, Z.; Gong, B.; Zhang, Y.; Shen, Y.; Gao, C.; Sang, N. A Recipe for Scaling up Text-to-Video Generation with Text-free Videos. *arXiv preprint arXiv:2312.15770* **2023**.
60. Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. LaVie: High-Quality Video Generation with Cascaded Latent Diffusion Models. *arXiv preprint arXiv:2309.15103* **2023**.
61. Wang, W.; Yang, Y. VidProM: A Million-scale Real Prompt-Gallery Dataset for Text-to-Video Diffusion Models. In Proceedings of the Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track, 2024.
62. Sun, K.; Huang, K.; Liu, X.; Wu, Y.; Xu, Z.; Li, Z.; Liu, X. T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-Video Generation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2025.
63. Wang, J.; Duan, H.; Jia, Z.; Zhao, Y.; Yang, W.Y.; Zhang, Z.; Chen, Z.; Wang, J.; Xing, Y.; Zhai, G.; et al. LOVE: Benchmarking and Evaluating Text-to-Video Generation and Video-to-Text Interpretation. *arXiv preprint arXiv:2505.12098* **2025**. https://doi.org/10.48550/arXiv.2505.12098.
64. Wang, J.; Duan, H.; Jia, Z.; Zhao, Y.; Yang, W.Y.; Zhang, Z.; Chen, Z.; Wang, J.; Xing, Y.; Zhai, G.; et al. T2VBench: Benchmarking Temporal Dynamics for Text-to-Video Generation. *arXiv preprint arXiv:2505.12098* **2024**.
65. Wang, J.; Ma, A.; Cao, K.; Zheng, J.; Zhang, Z.; Feng, J.; Liu, S.; Ma, Y.; Cheng, B.; Leng, D.; et al. WISA: World Simulator Assistant for Physics-Aware Text-to-Video Generation. *arXiv preprint arXiv:2503.08153* **2025**.
66. Wei, Y.; Zhang, S.; Qing, Z.; Yuan, H.; Liu, Z.; Liu, Y.; Zhang, Y.; Zhou, J.; Shan, H. DreamVideo: Composing Your Dream Videos with Customized Subject and Motion. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2024, pp. 6537–6549.
67. Weng, W.; Feng, R.; Wang, Y.; Dai, Q.; Wang, C.; Yin, D.; Zhao, Z.; Qiu, K.; Bao, J.; Yuan, Y.; et al. ART•V: Auto-Regressive Text-to-Video Generation with Diffusion Models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2024.
68. Wu, J.Z.; Fang, G.; Wu, H.; Wang, X.; Ge, Y.; Cun, X.; Zhang, D.J.; Liu, J.W.; Gu, Y.; Zhao, R.; et al. Towards A Better Metric for Text-to-Video Generation. *arXiv preprint arXiv:2401.07781* **2024**.
69. Wu, W.; Li, Z.; Gu, Y.; Zhao, R.; He, Y.; Zhang, D.J.; Shou, M.Z.; Li, Y.; Gao, T.; Zhang, D. DragAnything: Motion Control for Anything Using Entity Representation. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2024.
70. Xie, D.; Xu, Z.; Hong, Y.; Tan, H.; Liu, D.; Liu, F.; Kaufman, A.; Zhou, Y. Progressive Autoregressive Video Diffusion Models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2025.

 doi:10.20944/preprints202512.1476.v1

71. Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Yu, W.; Liu, H.; Wang, X.; Wong, T.T.; Shan, Y. DynamiCrafter: Animating Open-Domain Images with Video Diffusion Priors. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2024. https://doi.org/10.48550/arXiv.2310.12190.
72. Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072* **2024**.
73. Yang, X.; Tan, Z.; Nie, X.; Li, H. IPO: Iterative Preference Optimization for Text-to-Video Generation. *arXiv preprint arXiv:2502.02088* **2025**.
74. Yin, T.; Zhang, Q.; Zhang, R.; Freeman, W.T.; Durand, F.; Shechtman, E.; Huang, X. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
75. Yin, Y.; Feng, Y.; Yang, Y.; Tang, Z.; Zhang, Z.; Yang, Z.; Jiao, B.; Chen, J.; Li, J.; Zhou, S.; et al. Step-Video-T2V Technical Report: The Practice, Challenges, and Future of Video Foundation Model. *arXiv preprint arXiv:2502.10248* **2025**.
76. Yuan, S.; Huang, J.; Xu, Y.; Liu, Y.; Zhang, S.; Shi, Y.; Zhu, R.; Cheng, X.; Luo, J.; Yuan, L. ChronoMagic-Bench: A Benchmark for Metamorphic Evaluation of Text-to-Time-lapse Video Generation. In Proceedings of the Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track, 2024.
77. Yuan, S.; Huang, J.; Shi, Y.; Xu, Y.; Zhu, R.; Lin, B.; Cheng, X.; Yuan, L.; Luo, J. MagicTime: Time-lapse Video Generation Models as Metamorphic Simulators. In Proceedings of the Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track, 2024.
78. Yuan, X.; Baek, J.; Xu, K.; Tov, O.; Fei, H. Inflation With Diffusion: Efficient Temporal Adaptation for Text-to-Video Super-Resolution. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, January 2024, pp. 489–496.
79. Yuan, S.; Huang, J.; He, X.; Ge, Y.; Shi, Y.; Chen, L.; Luo, J.; Yuan, L. Identity-Preserving Text-to-Video Generation by Frequency Decomposition. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
80. Zhang, G.; Zhang, T.; Niu, G.; Tan, Z.; Bai, Y.; Yang, Q. CAMEL: CAusal Motion Enhancement Tailored for Lifting Text-driven Video Editing. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12345–12354.
81. Zhang, R.; Li, W.; Chen, H.; Wang, Y.; Liu, J. Style-A-Video: Agile Diffusion for Arbitrary Text-Based Video Style Transfer. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2024. https://doi.org/10.1109/LSP.2024.3398538.
82. Zhang, Z.; Liao, J.; Li, M.; Dai, Z.; Qiu, B.; Zhu, S.; Qin, L.; Wang, W. Tora: Trajectory-oriented Diffusion Transformer for Video Generation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
83. Zhang, D.J.; Wu, J.Z.; Liu, J.W.; et al. Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation. *International Journal of Computer Vision* **2025**, *133*, 1879–1893. https://doi.org/10.1007/s11263-024-02271-9.
84. Zhao, R.; Gu, Y.; Wu, J.Z.; Zhang, D.J.; Liu, J.W.; Wu, W.; Keppo, J.; Shou, M.Z. MotionDirector: Motion Customization of Text-to-Video Diffusion Models. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2024.
85. Zhou, Z.; Yang, Y.; Yang, Y.; He, T.; Peng, H.; Qiu, K.; Dai, Q.; Qiu, L.; Luo, C.; Liu, L. HiTVideo: Hierarchical Tokenizers for Enhancing Text-to-Video Generation with Autoregressive Large Language Models. *arXiv preprint arXiv:2503.11513* **2025**.
86. Zhu, Z.; Feng, X.; Chen, D.; Yuan, J.; Qiao, C.; Hua, G. Exploring Pre-trained Text-to-Video Diffusion Models for Referring Video Object Segmentation, 2024, [arXiv:cs.CV/2403.12042].
87. Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; et al. Improving Image Generation with Better Captions. https://cdn.openai.com/papers/dall-e-3.pdf, 2024.

 doi:10.20944/preprints202512.1476.v1

88. Esser, P.; Rombach, R.; Ommer, B. Taming Transformers for High-Resolution Image Synthesis. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12873–12883. https://doi.org/10.1109/CVPR46437.2021.01268.
89. Schuhmann, C.; Köpf, A.; Coombes, T.; Vencu, R.; Trom, B.; Beaumont, R. LAION-COCO: 600M synthetic captions from LAION2B-en. https://laion.ai/blog/laion-coco/, 2022. Accessed: 2025-08-06.
90. Ravi, N.; Gabeur, V.; Hu, Y.T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714* **2024**.