

Controllable Video Generation: A Survey

Yue Ma^{1†}, Kunyu Feng^{2†}, Zhongyuan Hu^{3†}, Xinyu Wang^{3†}, Yucheng Wang¹, Mingzhe Zheng¹, Bingyuan Wang², Qinghe Wang⁴, Xuanhua He¹, Hongfa Wang³, Chenyang Zhu³, Hongyu Liu¹, Yingqing He¹, Zeyu Wang^{1,2}, Zhifeng Li⁶, Xiu Li³, Sirui Han¹, Yike Guo¹, Wei Liu⁶, Dan Xu¹, Linfeng Zhang^{5*}, Qifeng Chen^{1*}

¹Hong Kong University of Science and Technology, Hong Kong, China.

²Hong Kong University of Science and Technology(Guang Zhou), Guang Zhou, China.

³Tsinghua University, Shenzhen, China.

⁴ Dalian University of Technology, Dalian, China.

⁵EPIC Lab of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai, China.

⁶Tencent, Shenzhen, China.

*Corresponding author(s). E-mail(s): zhanglinfeng@sjtu.edu.cn; cqf@ust.hk;

Contributing authors: mayuefighting@gmail.com; fengkunyu513@gmail.com;

huzhongyyuan@gmail.com; cpwxyxwcp@gmail.com; ywangls@connect.ust.hk;

mzhengar@connect.ust.hk; bwang667@connect.hkust-gz.edu.cn;

qinghewang@mail.dlut.edu.cn; xhecd@connect.ust.hk; wanghf22@mails.tsinghua.edu.cn;

chenyangzhu.cs@gmail.com; hliudq@connect.ust.hk; yhebm@connect.ust.hk;

zeyuwang@ust.hk; zhifeng0.li@gmail.com; li.xiu@sz.tsinghua.edu.cn; siruihan@ust.hk;

yikeguo@ust.hk; wl2223@columbia.edu; danxu@cse.ust.hk;

†These authors contributed equally to this work.

Abstract

With the rapid development of AI-generated content (AIGC), video generation has emerged as one of its most dynamic and impactful subfields. In particular, the advancement of video generation foundation models has led to growing demand for controllable video generation methods that can more accurately reflect user intent. Most existing foundation models are designed for text-to-video generation, where text prompts alone are often insufficient to express complex, multi-modal, and fine-grained user requirements. This limitation makes it challenging for users to generate videos with precise control using current models. To address this issue, recent research has explored the integration of additional non-textual conditions—such as camera motion, depth maps, and human pose—to extend pretrained video generation models and enable more controllable video synthesis. These approaches aim to enhance the flexibility and practical applicability of AIGC-driven video generation systems. In this survey, we provide a systematic review of controllable video generation, covering both theoretical foundations and recent advances in the field. We begin by introducing the key concepts and commonly used open-source video generation models. We then focus on control mechanisms in video diffusion models, analyzing how different types of conditions can be incorporated into the denoising process to guide generation. Finally, we categorize existing methods based on the types of control signals they leverage, including single-condition generation, multi-condition generation, and universal controllable

generation. For a complete list of the literature on controllable video generation reviewed, please visit our curated repository at <https://github.com/mayuelala/Awesome-Controllable-Video-Generation>.

Keywords: Video generation, Diffusion model, Controllable generation, Survey

1 Introduction

As interest in AI-generated content (AIGC) continues to grow, video generation, one of its key domains, has emerged as a prominent focus for both researchers and users alike. Modern video generation methods (Ho et al. 2022; Ma et al. 2024c; Chen et al. 2023b; Zhang et al. 2025j,g,c,b; Jiang et al. 2025a; Chen et al. 2025f; Zhang et al. 2025l; Feng et al. 2025a,b; Geng et al. 2025; Zhu et al. 2025a, 2024a; Long et al. 2025) typically leverage cutting-edge generative paradigms (*e.g.*, diffusion (Ho et al. 2020; Lipman et al. 2022; Wang et al. 2025e) or autoregressive models (Van den Oord et al. 2016; Zhang et al. 2025h,d,i; Ramesh et al. 2021; Chang et al. 2022; Zhang et al. 2025e; Tian et al. 2024a)), combined with large-scale datasets (Bain et al. 2021; Chen et al. 2024c; Schuhmann et al. 2021), massive model parameters (Kong et al. 2024; Ma et al. 2025b; Zhang et al. 2025f; Wan et al. 2025), and advanced architectural frameworks (Peebles and Xie 2023). We refer to these models as video generation foundation models, which have significantly advanced the quality of generated videos. The resulting outputs exhibit an unprecedented level of creativity. Despite their impressive generative capabilities, these models often remain constrained by their reliance on text-only conditioning, which limits the degree of control users can exert over the generated content. As a result, users frequently struggle to translate their creative ideas into precise video outputs, thereby diminishing the practical effectiveness of these models in real-world content creation scenarios.

To address this challenge, researchers have begun exploring ways to incorporate control signals beyond text, enabling more accurate and flexible guidance in the video generation process. For example, enabling users to modify camera trajectories or specify particular actions for characters in the video are emerging areas of interest. When fine-grained control over the generated content becomes possible, users are empowered with greater creative flexibility, thereby unlocking the

full potential and practical value of video generation as a task.

In this survey, we focus on the task of controllable video generation, including both its theoretical foundations and practical applications. Our goal is to *provide a comprehensive overview of the latest research advances and to shed light on the development trajectory of this rapidly evolving field*. Specifically, we start by providing a brief overview of the background and core concepts of video generative models, providing their theoretical basis. This analysis clarifies the core principles of earlier research, fostering a deeper understanding of the field. Subsequently, the detailed reviews of previous studies are conducted to emphasize their unique contributions and distinctive features. Then we investigate the wide-ranging applications of these methods, highlighting their practical significance and influence across various contexts and related downstream tasks. Additionally, we deep discuss the limitations and future work about controllable video generation. In Fig. 1, we present a line chart illustrating the number of controllable video generation studies utilizing various types of conditioning. As video foundation models have rapidly advanced, controllable video generation has also experienced significant growth.

Recent survey papers provide extensively overviews of AI-generated content (AIGC), covering various areas such as video generation based on Generative Adversarial Networks and Variational AutoEncoders (Aldausari et al. 2022; Yazdani et al. 2024), diffusion model theories and architectures (Xing et al. 2024d), efficient video diffusion models (Ma et al. 2025f), unified multi-modal video synthesis and understanding (Zhou et al. 2024a), video editing (Sun et al. 2024), foundational video diffusion models (Wang et al. 2025h; Yang et al. 2025b; Melnik et al. 2024), and 4D generation applications (Ren et al. 2025). While these reviews offer valuable insights, many only provide a cursory examination of video generative models

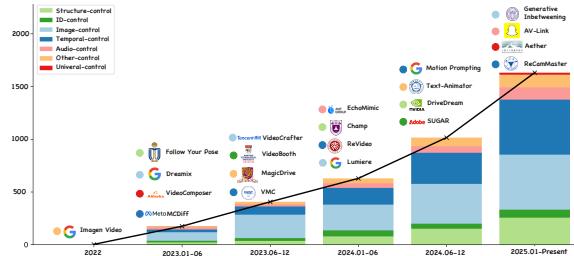


Fig. 1: The development trend of controllable video generation methods across seven representative task categories. The line chart illustrates the rapid growth in the number of related works from 2022 to the present, with different categories distinguished by color. Representative works from each period are highlighted above the chart. For instance, VideoCrafter (Chen et al. 2023b) and EchoMimic (Chen et al. 2025f) have achieved 4.9k and 3.9k stars in Github, respectively.

or predominantly concentrate on other modalities. This is a significant gap in the literature regarding controllable video generation. Additionally, existing studies rarely address this topic in various control signals, *e.g.*, depth, sketch, leaving a critical void in understanding the potential for integrating novel conditions into video generative models and their implications for advancing controllable video generation.

In summary, our contributions are as follows:

- A well-structured taxonomy of controllable video generation methods is presented by classifying existing methods according to their input control signals, which facilitates understanding of existing methods and reveals core challenges in this field.
- We present the theoretical foundations of GAN-, VAE-, Flow-, DM-, and AR-based architectures, along with recent video generation models built upon them, providing a clearer understanding of their underlying mechanisms.
- Our survey introduces broad coverage of conditional generation approaches, structured around the proposed taxonomy, and emphasizes the defining traits and methodological innovations of each technique.
- We investigate the practical impact of conditional generation within video models, covering a range of generative scenarios that reflect its increasing significance in the AIGC landscape. In addition, we identify key shortcomings

of existing techniques and propose potential avenues for further exploration.

The remainder of this paper is organized as follows: Sec. 2 provides a concise overview of various generative paradigms. In Sec. 4, we introduce representative video generation models, and presents a comprehensive taxonomy for controllable video generation. In Sec. 5, we outline different control mechanisms, explain how novel conditions can be incorporated into video generation models, and summarize existing methods based on our proposed taxonomy. Sec. 6 highlights key application scenarios of controllable video generation. Lastly, in Sec. 7, we discuss several limitations of current research from both technical and practical perspectives, and propose promising directions for future work.

2 Preliminaries

This section first explains the basic theories of generative models. In Fig. 2, we present an illustration of GAN, VAE, Flow-based Models, Diffusion Models, and Autoregressive Models, then we give the taxonomy of the controllable video generation tasks.

GANs and VAEs. Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are two classical models for generative modeling. GAN (Goodfellow et al. 2014) is based on a minimax game between a generator G and a discriminator D . The generator tries to produce realistic samples from random noise $\mathbf{z} \sim p(\mathbf{z})$, while the discriminator attempts to distinguish between real data $\mathbf{x} \sim p(\mathbf{x})$ and generated samples $G(\mathbf{z})$. The objective is given by

$$\min_G \max_D \mathbb{E}_{\mathbf{x}} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}} [\log (1 - D(G(\mathbf{z})))].$$

VAE (Kingma et al. 2013) introduces a variational posterior $q(\mathbf{z}|\mathbf{x})$ to approximate the true posterior $p(\mathbf{z}|\mathbf{x})$. The training objective is to maximize the Evidence Lower Bound (ELBO):

$$\begin{aligned} \mathcal{L}(\theta; \mathbf{x}) = & \mathbb{E}_{q(\mathbf{z}|\mathbf{x}; \theta)} [\log p(\mathbf{x} | \mathbf{z}; \theta)] \\ & - D_{\text{KL}}(q(\mathbf{z} | \mathbf{x}; \theta) \| p(\mathbf{z})). \end{aligned}$$

Both models form the foundation for developments in generative modeling, including diffusion and flow-based models.

Diffusion Models. In recent years, diffusion models have emerged as a powerful series of generative models, offering high-quality sample generation. A typical model is the Denoising Diffusion Probabilistic Model (DDPM) (Ho et al. 2020), which add Gaussian noise to pure data and learn to reverse this process at sample generation, following the Markov Chain.

- **Forward Process.** In the forward process, DDPM converts clean data from the previous data distribution to noise by gradually adding random Gaussian noise:

$$q(x_1, \dots, x_T | x_0) = \prod_{t=1}^T q(x_t | x_{t-1}),$$

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

where the noise schedule β_t is designed to increase monotonically with t , ensuring noise is smoothly added into the clean data.

- **Reverse Process.** In the reverse process, the models aim to learn the ability to restore data from noise through denoising x_t . At each step, the denoising operation is formulated as

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^T p_\theta(x_{t-1} | x_t),$$

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)).$$

where the mean $\mu_\theta(x_t, t)$ and the variance $\Sigma_\theta(x_t, t)$ are parameterized by neural networks trained to predict the denoising transformation from x_t to x_{t-1} . To simplify the reverse process, the diffusion models often reparameterize the mean $\mu_\theta(x_t, t)$ using a noise prediction model $\epsilon_\theta(x_t, t)$, which directly estimates the noise added at each timestep instead of recovering the clean data:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right).$$

Flow-based Models. Flow-based models are a class of powerful generative models, with the core idea being to learn an accurate, invertible

transformation that converts a simple base distribution into a complex target data distribution. This transformation process is typically composed of a series of invertible functions, referred to as “flows”.

Traditional flow-based models (Dinh et al. 2014; Chen et al. 2024b; Dinh et al. 2016; Kingma and Dhariwal 2018; Zhang et al. 2025m; Song 2022; Song et al. 2023; Song and Zhang 2022; Chen et al. 2025d) rely on discrete sequences of invertible transformations. In contrast, Continuous Normalizing Flows (CNFs) (Chen et al. 2018) generalize this framework by modeling the transformation as a continuous-time flow governed by an Ordinary Differential Equation (ODE):

$$\frac{d}{dt} \phi_t(x) = v_t(\phi_t(x)), \quad \phi_0(x) = x,$$

where ϕ_t is the flow map, describing the transformation of data from x_0 to x_t , and v_t is a time-dependent vector field, typically parameterized by neural networks, that defines the direction and magnitude of change.

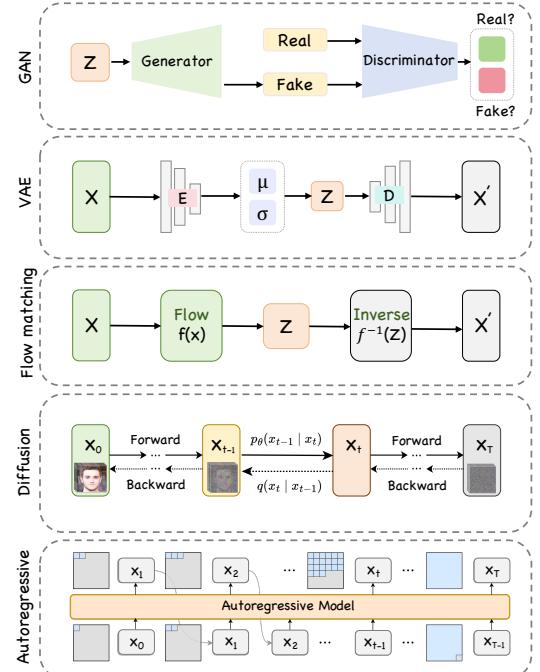


Fig. 2: Illustration of various generative methods. We show the method of GAN, VAE, Flow Matching, Diffusion Models, and Autoregressive Models for video generation.

To improve the efficiency of training CNFs, a modern generative approach called *Flow Matching* (FM) (Lipman et al. 2022) has been proposed. It offers a new paradigm for learning complex data distribution by directly learning the vector field that transfers samples from the base to the target distribution.

To learn the vector field v_t , Flow Matching introduces a novel supervision signal based on sample pairs (x_0, x_1) drawn from the base and target distributions, respectively. It defines a straight-line interpolation between the two:

$$x_t = (1 - t)x_0 + tx_1, v_t^{\text{gt}}(x_t) = x_1 - x_0,$$

The model is then trained by minimizing the squared error between the predicted and ground-truth velocity:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{x_0 \sim p_0, x_1 \sim p_1, t \sim u[0,1]} \left[\|v_t(x_t) - (x_1 - x_0)\|^2 \right].$$

This loss encourages the vector field to match the direction and magnitude of the straight-line transport from x_0 to x_1 .

Autoregressive Models. Autoregressive (AR) (Ramesh et al. 2021; Wu et al. 2021; Ding et al. 2021; Yu et al. 2022; Gafni et al. 2022; Chang et al. 2022; Tian et al. 2024a) Models generate data by modeling the conditional distribution of each element given the previous ones. Formally, given a sequence $\mathbf{x} = (x_1, x_2, \dots, x_T)$, the joint probability can be factorized as

$$P(\mathbf{x}) = \prod_{t=1}^T P(x_{t+1} | x_{\leq t}).$$

In the early stage, AR models like Pixel-CNN (Van den Oord et al. 2016) generate pixel row by row, to capture local dependencies. While, due to their inherently sequential generation process, suffering from efficiency, they were gradually replaced by diffusion models. Recently, with the growth of video generation and the strong temporal correlations between frames, AR models have regained attention. Many works adopt hybrid designs, integrating autoregressive priors with diffusion backbones.

3 Method

In this section, we describe the methodology for our literature collection and screening process. We focused on papers published within the past five years (2020–2025), as this period marks the rapid evolution of controllable video generation techniques driven by diffusion models. As illustrated in Fig. 3, we followed the PRISMA (Moher et al. 2009) guideline with a four-phase procedure to select and review relevant works. To align with our research scope, we established the following three inclusion criteria: **C1. Controllable Video Generation:**

The paper must present methods, frameworks, or systems that enable fine-grained control over video synthesis, beyond standard text-to-video generation. Acceptable control signals include but are not limited to: pose, depth, sketches, bounding boxes, motion trajectories, camera parameters, audio, identity, or multi-modal conditions. **C2. Based on Generative Foundation Models:** The paper should build upon widely recognized generative paradigms, such as DMs, ARs, GANs, flows, or hybrid architectures. **C3. Relevant Contributions:** We also include papers that introduce datasets, evaluation benchmarks, systematic analyses, or user-centric tools that facilitate controllable video generation, even if they do not propose new generative models.

Phase 1: Identification. To identify influential works in controllable video generation, we selected a series of publication venues based on their impact in computer vision and graphics: Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), Neural Information Processing Systems (NeurIPS), International Conference on Learning Representations (ICLR), International Conference on Machine Learning (ICML), ACM Multimedia (MM), Winter Conference on Applications of Computer Vision (WACV), International Conference on Acoustics, Speech, and Signal Processing (ICASSP), International Conference on Robotics and Automation (ICRA), SIGGRAPH/SIGGRAPH Asia, ACM Transactions on Graphics (TOG), IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Transactions on Visualization and Computer Graphics (TVCG), Transactions on Multimedia (TMM), and Transactions on

Machine Learning Research (TMLR). We also considered high-impact preprints on arXiv to capture the most recent advances. We constructed a set of search queries using keywords reflecting the two core aspects of our survey: (1) *controllability* and (2) *video generation*. Example queries used in ACM Digital Library and arXiv search are:

Title/Abstract: ((controllable video)
OR (controlled video generation) OR
(conditional video synthesis) OR (video control) OR (motion control) OR (camera control) OR (pose-guided video) OR (audio-driven video) OR (text-to-video) OR (image-to-video)) AND (generative model OR diffusion OR VAE OR GAN OR AR).

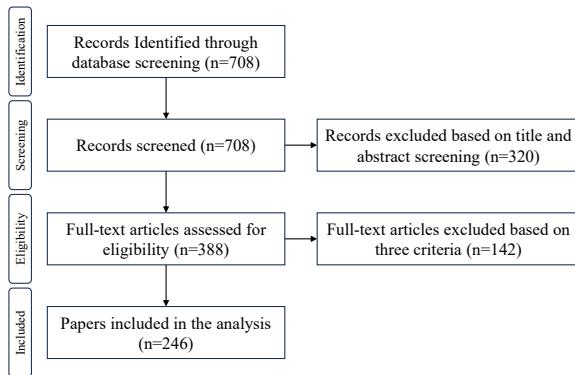


Fig. 3: PRISMA (Moher et al. 2009) flow diagram of our literature review process. We performed a four-phase screening covering identification, screening, eligibility assessment, and final inclusion, as detailed in Sec. 3. The number of papers retained at each stage is denoted by N.

We included papers published between January 2020 and October 2025, written in English. The initial search yielded n=708 candidate papers.

Phase 2: Screening. We screened the titles and abstracts of the 708 papers against criteria **C1–C3**. Papers that clearly did not address controllable video generation (*e.g.*, focusing solely on video recognition, compression, or non-generative tasks) were excluded. Additionally, papers that introduced generative models without any mechanism for external control were filtered out. After this phase, 388 papers remained for full-text

review. To ensure consistency, two authors independently screened a random subset of 100 papers; the inter-rater agreement measured by Cohen’s Kappa was 0.87.

Phase 3: Eligibility. We assessed the full text of the 388 papers for eligibility. Exclusion at this stage was based on: (a) failure to meet both **C1** and **C2** upon detailed reading, or (b) limited relevance to the core theme of controllable video generation despite meeting **C1** or **C2** superficially. For instance, papers that only briefly mentioned controllability as a future direction without substantial technical contribution were excluded. This phase removed 142 papers, leaving 246 papers for final inclusion.

Phase 4: Taxonomy. To provide a comprehensive and structured overview of the field, we propose a multi-layered framework that captures the entire ecosystem of controllable video generation, spanning from technical foundations to practical applications. As illustrated in Fig. 4, our framework encompasses six interconnected dimensions: **Technical Foundations:** The generative paradigms that form the basis of modern video synthesis, including GANs, VAEs, DMs, Flows, ARs, and corresponding foundation models. **Controllable Modules:** The specialized components designed to inject external conditions into pre-trained generative models, such as ControlNet, adapters, encoders, and other modules. **Control Modalities:** The types of input signals used to guide video generation, categorized into seven distinct classes based on their semantic and structural characteristics. **Creative Support:** The user-centric capabilities that empower creators through iterative design, real-time preview and adjustment, integrated control, etc. **Technical Tasks:** The core computer vision tasks enabled by controllable generation, including video completion and inpainting, video composition, video-to-4D generation, etc. **Practical Applications:** The real-world domains where these technologies demonstrate significant impact, such as film production, autonomous driving, and digital humans.

Within this comprehensive framework, we adopt **control modality** as our primary taxonomic dimension for three reasons. First, the type of control signal fundamentally determines the architectural design choices, conditioning mechanisms, and technical challenges in video generation pipelines. Second, existing research efforts are

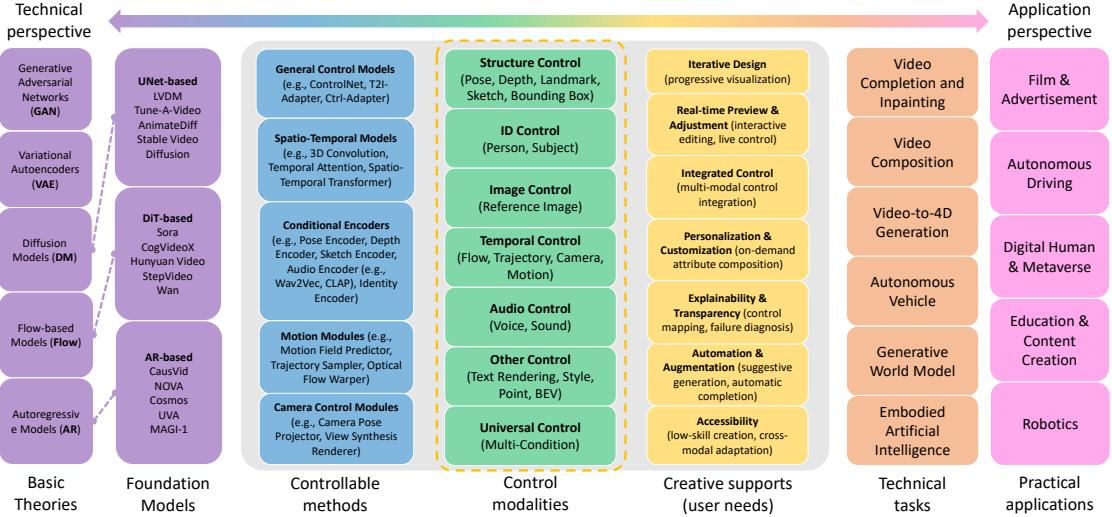


Fig. 4: An overall framework for controllable video generation. The seven control modalities form the primary classification axis. The framework connects basic theories (Sec. 2) and video generation foundation models (Sec. 4) with controllable methods and creative supports as key insights (Sec. 5), and maps them to downstream technical tasks and practical applications (Sec. 6).

predominantly organized around specific control types such as pose-guided animation or audio-driven synthesis, making this classification naturally align with the current literature structure. Third, control modalities serve as the critical bridge between user intent (expressed through various input forms) and the generated visual content. As shown in Fig. 4, we systematically categorize controllable video generation methods into seven modality classes: (1) *Structure Control* (e.g., pose, depth, sketch, bounding boxes), (2) *ID Control* (person or subject identity), (3) *Image Control* (reference images), (4) *Temporal Control* (flow, trajectory, camera, motion), (5) *Audio Control* (voice and sound), (6) *Other Control* (text rendering, style, points, BEV), and (7) *Universal Control* (multi-condition integration).

The task of controllable generation with text-to-video diffusion models represents a multi-faceted and complex field. From the condition perspective, we divide this task into seven sub-tasks. Most works study how to generate video under specific conditions, e.g. pose-guided generation, and depth-guided generation. To reveal the underlying mechanisms and characteristics of these approaches, we further categorize them according to their condition types. The primary challenge in this field is enabling pretrained T2V

diffusion models to handle novel types of conditions and generate videos that are not only aligned with the provided textual prompts but also exhibit high visual quality and strong temporal consistency. Moreover, recent methods investigate how to generate videos using multiple conditions, such as given a reference image, sparse trajectory, and motion brush. The main challenge in these tasks is the effective integration of multiple, potentially interacting, conditions, necessitating the capability to express several conditions simultaneously and consistently throughout the generated video sequence.

4 Video Generation Foundation Models

Video generation foundation models have recently attracted substantial attention due to their remarkable ability to generate high-fidelity videos. These models are typically categorized into two major paradigms: diffusion-based models and autoregressive (AR) models. Among them, diffusion-based methods have shown impressive performance by modeling data distributions through an iterative denoising process, effectively capturing complex spatial-temporal dependencies.

Table 1: An overview of notable video generative models. Resolution: The frame count, width and height of generated video. f : downsampling factor of autoencoder from pixel space to latent space. \dagger : The method uses one-shot data from the Dataset.

Model	Venue	Param.	Resolution	f	Text Encoder	Training Dataset	Open Source	Latency
<i>UNet-Based Video Diffusion Models</i>								
IVDM (Ue et al. 2022)	arXiv: 2022	1:2B	Multi-Frame [103:1024] \times 256 2	4, 8x8	CLIP-ViT-L/14	WebVis-2M (Bain et al. 2021)	✓	114
Time-A-Video (Wu et al. 2023)	ICCV: 2023	983M	24 \times 512 2	1, 8x8	CLIP-ViT-L/14	DAVIS [†] (Pont-Tuset et al. 2017)	✓	769
AnimatedDiff (Gao et al. 2024c)	ICLR: 2024	1.4B	16 \times 256 2	1, 8x8	CLIP-ViT-L/14	WebVis-10M (Bain et al. 2021)	✓	122
Stable Video Diffusion (Blattmann et al. 2023)	arXiv: 2023	1.5B	14 \times 576 \times 1024	1, 8x8	CLIP-ViT-H/14	Internal Dataset	✓	98
<i>DiT-Based Video Diffusion Models</i>								
CogVideoX (Yang et al. 2025d)	ICLR: 2025	2B3:5B	49 \times 480 \times 720	4, 8x8	T5	Internal Dataset	✓	180
HunyuanVideo (Kong et al. 2024)	arXiv: 2024	13B	129 \times 720 \times 1280	4, 8x8	Hunyuan-Large	Internal Dataset	✓	3448
StepVideo (Ma et al. 2025b)	arXiv: 2025	30B	204 \times 544 \times 992	8, 16x16	Hunyuan-CLIP, Step-LLM	Internal Dataset	✓	743
Wan (Wan et al. 2025)	arXiv: 2025	1.3B3:14B	81 \times 720 \times 1280	4, 8x8	uniT5	Internal Dataset	✓	1864
<i>Video Autoregressive Models</i>								
CausVid (Yin et al. 2025)	CVPR: 2025	1.4B	120 \times 352 \times 640	4, 8x8	uniT5	Internal Dataset	✓	14
NOVA (Dong et al. 2024)	ICLR: 2025	0.3B3:0.4B3:1.4B	33 \times 352 \times 640	4, 8x8	P16x8	Internal Dataset	✓	354
CommonAvatar (Li et al. 2025)	arXiv: 2025	4B3:12B	121 \times 768 \times 1280	4, 8x8	T5-XXL	Internal Dataset	✓	119
UVA (Li et al. 2025c)	RSS: 2025	0.5B	16 \times 720 \times 1280	1, 16x16	CLIP-ViT-B/32	Libero10 (Lin et al. 2023a), PushT (Chi et al. 2024a), UMI (Chi et al. 2024b), Human Video (Clark et al. 2025)	✓	j
MAGI-1 (Teng et al. 2025)	arXiv: 2025	4.5B3:24B	24 \times 720 \times 1280	4, 8x8	T5	Internal Dataset	✓	1740
InfinityStar (Lin et al. 2025b)	arXiv: 2025	8B	81 \times 720 \times 1280	4, 8x8	T5	Internal Dataset	✓	58

We introduce the typical frameworks among them (see in Table 1). Diffusion-based video generation models can be further divided into two architectural branches: those built on UNet-based frameworks and those leveraging the more recent DiT (Diffusion Transformer) architecture. We provide them in supplementary material.

5 Controllable Video Generation Models with Various Conditions

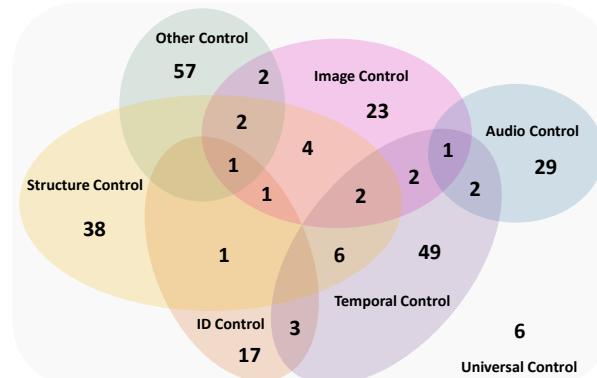


Fig. 5: Venn diagram of controllable video generation literature categorized by control modalities. The diagram illustrates the distribution of surveyed works across seven control categories, including Structure, Image, Audio, Temporal, ID, Universal, and Other controls. The overlap regions highlight methods that support multiple control signals simultaneously. The counts in each region indicate the number of papers falling into the corresponding category intersections, reflecting current research trends and the degree of multimodal controllability explored in existing video generation models.

This section forms the core of our survey, delving into the diverse methodologies developed for controllable video generation. As shown in Fig. 6 and Table. 2, we present various visual illustrations and representative works of controllable video generation. These approaches are categorized on the basis of the primary nature of the control signal used to guide the synthesis process. Specifically, we will explore methods focusing on structure control, ID control, image control, temporal control, audio control, other control, and universal control. For each category, we will review seminal and recent works, highlighting their foundational techniques, architectural innovations, and the specific challenges.

5.1 Structure Control

Structure control in video generation refers to the ability to dictate the spatial layout, conformation of articulated objects (like humans or animals), and the geometric properties of the scene. This form of control is paramount for producing videos that are not only visually coherent but also adhere to plausible physical configurations and user-specified arrangements. By conditioning the generation process on structural cues, models can synthesize complex scenes with greater fidelity and semantic correctness.

5.1.1 Pose-Guided Generation

Pose-guided video generation specifically aims to animate subjects, predominantly humans, according to a sequence of predefined poses. This is a critical task for applications such as virtual avatar animation, character generation for interactive entertainment, human motion synthesis from sparse inputs, and fashion video synthesis (Karras et al. 2023; Jin et al. 2024; Wang et al. 2025i;

Gong et al. 2025; Ma et al. 2025c,d; Wang et al. 2024c). The core challenge lies in generating temporally coherent video frames where the subject’s appearance is preserved while accurately following the target pose sequence $P = p_1, p_2, \dots, p_T$, where each p_t represents the pose in frame t .

The input pose information p_t is typically represented as 2D skeletal keypoints (e.g., COCO format), 3D skeletal coordinates, or more dense representations like DensePose (Güler et al. 2018), which maps image pixels to 3D surface coordinates of the human body. These pose sequences can be extracted from existing videos using off-the-shelf pose estimators (e.g., OpenPose (Cao et al. 2019)), derived from motion capture (MoCap) data, or generated algorithmically. The choice of pose representation influences the granularity of control and the complexity of the generation task.

Numerous approaches have been proposed to achieve pose-guided generation. Early methods (Yang et al. 2018) often rely on GANs, using pose information to condition the generator, for instance, by rendering pose stick figures and feeding them as input alongside a latent code. More recent work leverages the power of diffusion models. For example, MagicAnimate (Xu et al. 2024e) and Animate Anyone (Hu 2024) and its successor Animate Anyone 2 (Hu et al. 2025a) utilize diffusion models conditioned on pose sequences and a reference image to generate temporally consistent animations. Techniques like Champ (Zhu et al. 2024b) and MimicMotion (Zhang et al. 2024c) focus on high-fidelity human motion synthesis. DirectorLLM (Song et al. 2025b) employs large language models to direct human-centric video generation based on pose and textual descriptions. Follow Your Pose (Ma et al. 2024c) introduces a two-stage training strategy to create high-quality character videos. ControlNet-based approaches (Zhang et al. 2023a) have also been adapted, where pose maps serve as direct spatial conditions for pre-trained text-to-image diffusion models, which are then fine-tuned or augmented with temporal modules for video tasks (Peng et al. 2024; Wang et al. 2024e). Disentangling pose from appearance is a key strategy, often achieved by dedicated appearance encoders and pose encoders, with mechanisms like feature warping or attention to align appearance features with the target pose (Li et al. 2024a). Temporal consistency is

often enforced through temporal attention mechanisms across frames or by incorporating temporal smoothness losses.

Despite significant progress, pose-guided generation still faces challenges. Handling severe self-occlusions, where parts of the subject are hidden due to the pose, requires robust inpainting capabilities. Maintaining fine-grained texture details and consistent identity across large pose variations remains difficult, especially when appearance must be reliably transferred from a reference image without explicit identity control. Ensuring natural and smooth transitions between poses, avoiding jittery or robotic movements, is crucial for realism. Furthermore, generalization to unseen poses, body shapes, or complex clothing items not well-represented in training data can be problematic, often requiring extensive and diverse datasets like those proposed by FCGV (Zhu et al. 2025b), which aims for broader character controllability.

5.1.2 Depth-Guided Generation

Depth-guided video generation utilizes depth maps as conditioning signals to control the 3D structure of the synthesized scene, object layout, and relative distances of elements from the camera. This form of control is crucial for generating videos with a strong sense of three-dimensionality, realistic object interactions, and consistent scene geometry across frames.

The primary input for depth-guided control is a sequence of depth maps $D = d_1, d_2, \dots, d_T$, where each $d_t \in \mathbb{R}^{H \times W}$ provides per-pixel depth information. These depth maps can be relative or absolute, estimated from RGB videos using monocular depth estimation models, captured using specialized sensors (e.g., LiDAR, ToF cameras), or rendered from 3D models. Depth conditioning helps in maintaining geometric consistency, enabling plausible occlusions, and can be used to simulate effects like depth-of-field.

Several works have explored depth-guided video synthesis. ControlVideo (Zhang et al. 2023b) and Control-A-Video (Chen et al. 2023d) demonstrate how various control signals, including depth, could be integrated into large pre-trained text-to-image diffusion models to guide video generation. SparseCtrl (Guo et al. 2024d) focuses on using sparse depth controls. DreamDance (Pang et al. 2024) leverages depth for dynamic 3D scene

Table 2: Representative works of controllable generation with video generative model.

Type	Method	Venue	Model	Condition	Training Dataset
Structure Control	Follow-Your-Pose Ma et al. (2024a)	AAAI 2024	UNet	Pose, Landmark, Text	LAION-400M Schuhmann et al. (2021) , HDVILA Xue et al. (2021) WebVid-10M Bain et al. (2021) Self-Construction Datasets muScenes Cesaer et al. (2019) HDFP Zhang et al. (2021) , CelebV-HQ Zhu et al. (2022)
	Make-Your-Video Xing et al. (2024b)	TVCG 2024	UNet	Depth, Text	
	ToonCrafter Xing et al. (2024c)	TOG 2024	UNet	sketch, Image, Text	
ID Control	DriveDreamer Zhao et al. (2025a)	AAAI 2024	UNet	BBox, Image, Text	WebVid-10M Bain et al. (2021) , LAION-400M Schuhmann et al. (2021) Panda-70M Chen et al. (2024c) , Subject200k Tao et al. (2025) , OmniGen Xiao et al. (2025a)
	EchoMimic Meng et al. (2025a)	AAAI 2024	UNet	Audio, Landmarks, Image	
Image Control	VideoBooth Jiang et al. (2024b)	CVPR 2023	UNet	Object, Text	WebVid-10M Bain et al. (2021) , LAION-COCO-600M LAI (2025) , WebVid-103M Bain et al. (2021) Self-Construction Datasets WebVid-10M Bain et al. (2021) WebVid-10M Bain et al. (2021)
	Vlogger Zhang et al. (2024)	CVPR 2024	UNet	Object, Text	
	Phantom Lin et al. (2025d)	ICCV 2025	DiT	Subject, Text	
Temporal Control	VideoCrafter1 Chen et al. (2023b)	arXiv 2023	UNet	Image, Text	LAION-COCO-600M LAI (2025) , WebVid-103M Bain et al. (2021) Panda-70M Chen et al. (2024c) , Peacock Pex (2025)
	LunarBar-Tai et al. (2024)	SIGGRAPH-Asia 2024	UNet	Image, Mask, Style, Text	
	Concert-Edit Chen et al. (2024a)	TOG 2024	UNet	Image, Text	
Audio Control	Follow-Your-Click Lin et al. (2025e)	AAAI 2025	UNet	Click, Image, Text	RealEstate10K Zhou et al. (2018) , DL3DV Ling et al. (2023) MovieShot Rao et al. (2020)
	NOVA Deng et al. (2024)	ICLR 2025	Autoregressive	Image, Text	
	Motion-3DV Shi et al. (2024)	SIGGRAPH 2024	UNet	Trajectory, Camera, Text	
Other Control	MotionBooth Wu et al. (2024a)	NeurIPS 2024	UNet	Motion, Camera, BBox, Image, Text	WebVid-10M Bain et al. (2021) Panda-70M Chen et al. (2024c) HumanML3D Guo et al. (2022) , BEAT2 Lin et al. (2023b) , FineDance Li et al. (2023)
	ViewCrafter Wu et al. (2024a)	TPAMI 2025	UNet	Camera, Image	
	Direct-A-Video Yang et al. (2024b)	SIGGRAPH 2024	UNet	Camera, Text	
Universal Control	Panacea Wen et al. (2024)	CVPR 2023	UNet	BEV, Text	muScenes Cesaer et al. (2019) Self-Construction Datasets None WebVid-10M Bain et al. (2021)
	StyleMaster Ye et al. (2025)	CVPR 2024	DiT	Style, Video, Text	
	UniVST Song et al. (2024a)	TPAMI 2025	DiT	Style, Mask, Video	
Universal Control	GS-DiT Bian et al. (2025a)	CVPR 2025	DiT	Point, Video, Text	MiraData Ju et al. (2024) , RealEstate10K Zhou et al. (2018) , ConceptMaster Huang et al. (2025c) , Panda-70M Chen et al. (2024c) Self-Construction Datasets
	VideoComposer Wang et al. (2023a)	NeurIPS 2023	UNet	Universal Conditions	
	FullDiT Ju et al. (2025)	ICCV 2025	DiT	Universal Conditions	
	VACE Jiang et al. (2025b)	ICCV 2025	DiT	Universal Conditions	

generation. ControlNeXt ([Peng et al. 2024](#)) and Champ ([Zhu et al. 2024b](#)) also support depth guidance among other control types. Depth information is typically encoded and injected into the diffusion models’ U-Net architecture, often concatenated with noisy latents or fed into cross-attention layers. Some methods may also use depth-based loss functions to further encourage geometric fidelity, for instance, by comparing the depth map of a generated frame with the target depth map.

Challenges in depth-guided generation include the difficulty of obtaining accurate, dense, and temporally consistent depth maps, especially from monocular RGB videos. Errors or noise in the input depth maps can propagate to the generated video, leading to distorted geometry. Ensuring that the generated RGB frames are consistent with the conditioning depth information is non-trivial. Dynamic scenes with rapidly changing depths or complex non-rigid object deformations pose further difficulties. Moreover, incorporating 3D awareness often increases computational complexity. Approaches like Make-Your-Video ([Xing et al. 2024b](#)) and Moonshot ([Zhang et al. 2024a](#)) aim to improve fidelity and consistency in such scenarios.

5.1.3 Landmark-Guided Generation

Landmark-guided control focuses on using key-points or landmarks (*e.g.*, facial features, body joints, or other salient points) to guide video synthesis. These landmarks serve as spatial constraints that define the motion or deformation of specific regions, enabling fine-grained control over facial expressions, gestures, or body poses. This

approach is particularly valuable for applications like portrait animation, emotion-driven synthesis, and pose-guided video generation, where subtle movements and precise alignment with landmarks are critical for achieving natural and expressive results.

Landmark control signals can be provided as static or dynamic landmark sequences extracted from images or videos, or as high-level descriptions that are converted into landmark trajectories. For instance, TASTE-Rob ([Zhao et al. 2025b](#)) uses temporal landmark sequences to guide video generation. Follow-Your-Pose ([Ma et al. 2024c](#)) and Follow-Your-Emoji ([Ma et al. 2024d](#)) allow pose-based and emoji-driven landmark control, respectively, offering users intuitive ways to define movements and expressions.

Recent landmark-guided methods leverage landmarks to condition generative models, ensuring spatial alignment and temporal coherence. HunyuanPortrait ([Xu et al. 2025](#)) introduces implicit conditioning mechanisms to generate highly realistic and controllable portrait animations. Takin-ADA ([Lin et al. 2024b](#)) and EchoMimic ([Chen et al. 2025f](#)) specialize in audio-driven portrait animation, aligning facial landmarks with speech dynamics. EchoMimicV2 ([Meng et al. 2025a](#)) extends this by simplifying body and facial control for efficient synthesis. LivePortrait ([Guo et al. 2024a](#)) and IPT2V ([Yuan et al. 2025](#)) emphasize identity preservation while animating portraits, ensuring that synthesized videos remain faithful to the input identity.

Landmark-guided methods face significant challenges in maintaining accurate spatial alignment, especially when landmarks are sparse, noisy,

or incomplete. Ensuring smooth and natural transitions between frames while adhering to landmark constraints is critical for avoiding artifacts and achieving realistic motion. Capturing subtle expressions or complex dynamics, such as emotion or speech-driven facial movements, can be difficult without introducing distortions or losing fidelity.

5.1.4 Sketch-Guided Generation

Sketch-guided video generation utilizes single sketch frames or combines them with reference visual images to improve the performance of traditional text-to-video diffusion models. Sketch input becomes particularly valuable in applications such as quick character animation, where users can sketch the key poses or actions to guide the video generation process, and scene composition, where simple sketches can define object placement and movement in a visual story. This approach is especially beneficial for colorization applications like storytelling, allowing creators to quickly visualize and refine their ideas before detailed production.

In previous works, works like SketchBetween (Loftsdottir and Guzdial 2022) introduce a VAE-based framework for the sketch-based video generation application. Recently, the ControlNet structure (Zhang et al. 2023a) has been applied to sketch-guided video generation, but directly using it does not ensure the frame-to-frame consistency essential for coherent video production. As the existing control method commonly relies on dense temporal maps, SparseCtrl (Guo et al. 2024d) proposed an auxiliary encoder like ControlNet, which utilizes the sparse condition maps to generate the target video. Similarly, CTRL-Adapter (Lin et al. 2024c) designed a control adapter that supports various inputs, including image, depth, and sketch, facilitating the performance of controllable video generation applications. OpenSoraPlan (Lin et al. 2024a) proposed a powerful video generation framework with a condition encoder.

Numerous works also focus on exploring sketch-guided video generation applications, such as animation, video colorization, and many other tasks. TaleCrafter (Gong et al. 2023), a visual storytelling system, designs a controllable text-to-image component that generates target images based on input conditions and converts them into video through a subsequent Image-to-Video (I2V) module. ToonCrafter (Xing et al. 2024c)

proposes a cartoon video interpolation method, which designs a sketch encoder to provide efficient drawing tools. AniDoc (Meng et al. 2025b) and LVCD (Huang et al. 2024) also focus on the sketch-guided animation application. For general video generation, methods like MakePixels-Dance (Zeng et al. 2024), VideoLCM (Wang et al. 2023b), and TF-T2V (Wang et al. 2024g), these methods not only improve the quality and the controllability of the video but also optimize the consistency for the long-time video generation task. Besides, MagicStick (Ma et al. 2025c), SketchVideo (Liu et al. 2025a), MotionClone (Ling et al. 2025), and CameraCtrl (He et al. 2024a) utilize the sketch-guided controllable component for video editing and camera motion video generation.

5.1.5 BBox-Guided Generation

BBox-guided video generation directs the synthesis of video content by using bounding boxes as precise conditional inputs. This guidance can range from a static layout defining object placement in a single frame to a sequence of moving boxes dictating an object’s trajectory over time. The technique is prominently explored in autonomous driving video models, where bounding boxes are used to control the overall layout of the environment, as well as the precise position and movement of vehicles within the scene.

To inject bbox information into generative models, researchers have developed various strategies, which can generally be categorized into two main types: cross-attention-based injection and additive encoder-based injection. Magic-Drive (Gao et al. 2023), aiming for fine-grained control over 3D geometry, pioneers the separate handling of different layout types. Delphi (Ma et al. 2024a) also injects layout embeddings via cross-attention. Another early work, Drive-Dreamer (Zhao et al. 2025a), seeking to build a real-world driving model, fuses the representations of 3D bounding boxes ($B_i \in \mathbb{R}^{N \times N_B \times 16}$) with visual features using a gated self-attention mechanism. To tackle the challenge of learning box-object correlations from purely visual cues, Boximator (Wang et al. 2024b) and Panacea (Wen et al. 2024) introduce the new self-attention layer into the existing U-Net blocks, which processes control tokens derived from Fourier-encoded box coordinates and object IDs. Ctrl-V (Luo et al.

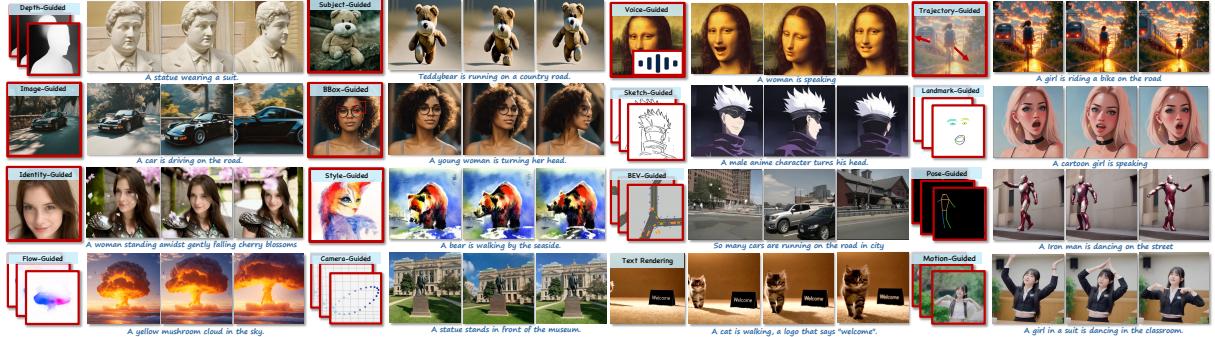


Fig. 6: Visual illustration of controllable video generation. We show the cases of controllable video generation with specific control. The conditions are marked in **red** and the prompts are marked in **blue**.

Table 3: Quantitative comparison of Landmark-guided video generation on 512×512 test images. LMD multiplied by 10^{-3} , AED multiplied by 10^{-2} , APD multiplied by 10^{-3} and Identity Similarity multiplied by 10^{-1} . \uparrow indicates higher is better. \downarrow indicates lower is better. The best results are in bold.

Method	Self Reenactment						Cross Reenactment					
	LMD \downarrow	FID-VID \downarrow	FVD \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	AED \downarrow	APD \downarrow	Identity \uparrow Similarity	Facial \uparrow Movements	Video \uparrow Quality	Temporal \uparrow Smoothness
LivePortrait (Guo et al. 2024a)	9.14	82.71	483.38	31.41	0.72	0.22	26.83	26.97	8.71	3.00	4.11	4.06
AniPortrait (Wei et al. 2024a)	6.67	81.90	430.24	30.54	0.67	0.27	22.42	21.32	7.95	3.83	4.00	3.81
FollowYE (Ma et al. 2024d)	5.63	77.15	417.53	30.84	0.68	0.25	20.06	20.95	8.18	4.11	4.23	3.33
X-Potrait (Xie et al. 2024)	6.23	82.93	416.41	30.81	0.71	0.19	20.66	20.39	8.03	3.88	3.73	3.49
HunyuanPortrait (Xu et al. 2025)	2.02	75.81	333.48	32.98	0.81	0.11	19.45	19.20	8.87	4.55	4.69	4.61

2025a) also utilizes a ControlNet architecture but introduces a unique two-stage pipeline. It first employs a diffusion-based BBox Generator to predict a complete video of the bounding box trajectories.

Moreover, some methods explore different injection methods. LLM-Grounded Video Diffusion (LVD) (Lian et al. 2023) takes a novel approach by leveraging a Large Language Model (LLM) to generate Dynamic Scene Layouts from text prompts. Also operating without training, TrailBlazer (Ma et al. 2024b) directly edits spatial and temporal attention maps during the initial denoising steps to guide subjects along key-framed bounding box trajectories. Recently, architectures based on the DiT have gained attention for their excellent scalability. For instance, DIVE (Jiang et al. 2024a), aiming to generate multi-view consistent videos, integrates a ControlNet-Transformer into the DiT architecture to process road information and uses a joint cross-attention mechanism to fuse scene and instance layouts.

5.1.6 Summary

In video generation, structural control acts as spatial cues, providing the foundational framework for dynamic visuals. Its core function is to precisely translate various abstract spatial requirements—from precise joint poses to casual user sketches—into natural and plausible visual movements. Early GAN-based frameworks (Yang et al. 2018) often conditioned the generation process by directly rendering structural skeletons together with latent features. And this paradigm was subsequently dominated by UNet-based architectures which employs Adapter interfaces (*e.g.*, ControlNet (Zhang et al. 2023a)) to graft spatial constraints onto main network. Currently, the research focus has clearly shifted towards diffusion Transformers(*e.g.*, MagicDrive-DiT (Gao et al. 2025)) due to the scaling limits of CNNs.

This shift is from the user’s need for finer control granularity and temporal coherence, with trends leaning toward integrated approaches that enable combining multiple structural signals (*e.g.*, pose with depth in DreamDance (Pang et al. 2024)). Beyond interaction, profound technical

challenges persist in physical awareness and generalization. Current methods often prioritize visual alignment over kinematic plausibility, leading to physically impossible deformations, and remain limited to specific domains (*e.g.*, humans). Consequently, future improvement is needed for interaction simplicity and physics-aware world modeling. The ultimate goal is to bridge the gap between abstract control and concrete realization, ensuring outcomes are not only physically consistent but also faithfully aligned with the user’s creative intent.

5.2 ID Control

ID control in video generation refers to the ability to preserve and manipulate the visual identity of specific entities within the whole video generation, such as humans, animals, or objects. This form of control is essential for ensuring temporal consistency in appearance, enabling personalized or instance-specific content generation. By conditioning on reference images or identity embeddings, models can maintain subject integrity across varying poses, motions, and viewpoints.

5.2.1 Person-Guided Generation

Person-guided video generation aims to synthesize videos of a specific individual while maintaining high-fidelity identity consistency throughout the video. It enables personalized animation where a character—typically provided as one or more reference images—is animated according to user-defined prompts or motion signals. This form of control is essential for digital avatars, identity-preserving storytelling, and interactive applications, where both visual likeness and behavioral alignment must be achieved.

The primary input for person-guided control includes one or more ID images of the target person. These can be facial portraits or full Face and body synthesis. A driving signal—such as a text prompt, pose sequence, or audio—is also provided to guide motion. The ID images capture static visual identity, while the driving signal controls dynamics. Some methods, like ID-Animator(He et al. 2024b), use a single image, while others, like Movie Weaver(Liang et al. 2025) rely on a set of images covering both face and body.

Representative diffusion-based methods explore various architectures and identity-injection strategies. ID-Animator(He et al. 2024b) integrates a ViT-based face adapter via cross-attention, while Movie Weaver(Liang et al. 2025) leverages a DiT backbone with anchor-concept prompts for tuning-free generation. PersonalVideo(Li et al. 2025a) and Vlogger (Zhuang et al. 2024) apply a reward-based objective to a frozen LDM for identity-motion alignment. ConsisID (Yuan et al. 2025) and Magic-Me (Ma et al. 2025g) introduce frequency-aware and dynamic attention mechanisms.

5.2.2 Subject-Guided Generation

Subject-guided generation (Zhang et al. 2024d; Jiang et al. 2024b; Song et al. 2025c,d) refers to the process of generating video content where the primary focus is on guiding the generation with specific subjects, such as particular characters, objects, or scenes, to ensure that the generated video adheres to user-defined subject constraints. This method enables the customization of content by explicitly controlling which subjects appear, how they interact, and in what context, offering more precise control over the generated content. Initial research often concentrated on single-subject customization, as exemplified by VideoBooth (Jiang et al. 2024b), which utilizes image prompts for enhanced content control, and DreamVideo (Wei et al. 2024b), which decouples subject learning from motion learning. However, the field has rapidly progressed to address the more complex challenge of multi-subject video generation. Studies such as Disen-Studio (Chen et al. 2024a), CustomVideo (Wang et al. 2024i), and MovieWeaver (Liang et al. 2025) are at the forefront of addressing key difficulties in multi-subject scenarios. These challenges include ensuring the correct co-occurrence and temporally consistent appearance of multiple subjects, preserving their visual identities, avoiding attribute binding issues, and accurately assigning actions. Such approaches typically extend pretrained video diffusion models by incorporating several key innovations: novel fine-tuning strategies (*e.g.*, VideoDreamer’s (Chen et al. 2023a) Disen-Mix Finetuning, DisenStudio’s (Chen et al. 2024a) motion-preserved disentangled fine-tuning); advancements in attention

Table 4: Quantitative comparison of landmark-guided video generation on TED-talks
Dataset. L1 multiplied by 10^{-4} . \uparrow indicates higher is better. \downarrow indicates lower is better. The best and secondary results are in bold and underline, respectively. MRAA* and TPSMM* proposed these methods using real videos as driving signals. Therefore, they are not involved in quantitative comparison.

Method	FID \downarrow	SSIM \uparrow	PSNR \uparrow	LPIPS \downarrow	L1 \downarrow	FID-VID \downarrow	FVD \downarrow
MRAA*(Siarohin et al. 2021)	25.76	0.807	32.62	0.216	0.36	12.20	126.28
TPSMM*(Zhao and Zhang 2022)	24.69	0.819	32.92	0.205	0.35	9.04	87.76
MRAA (Siarohin et al. 2021)	50.36	<u>0.762</u>	<u>31.90</u>	0.266	<u>0.50</u>	<u>82.79</u>	493.02
TPSMM (Zhao and Zhang 2022)	23.71	0.771	32.30	0.252	<u>0.49</u>	32.12	260.67
DisCo (Wang et al. 2024e)	75.48	0.575	27.99	0.309	1.21	66.18	393.04
MagicAnimate (Xu et al. 2024e)	41.58	0.529	28.28	0.310	1.73	33.61	223.54
MagicPose (Chang et al. 2023)	<u>23.39</u>	0.723	30.08	<u>0.236</u>	0.81	<u>27.53</u>	<u>214.23</u>
FollowYourPosev2 (Xue et al. 2024)	18.21	0.779	30.88	0.198	0.46	10.24	81.73

mechanisms (*e.g.*, CustomVideo’s (Wang et al. 2024i) attention control with object segmentation); and the development of specialized datasets.

To further enhance the quality, controllability, and usability of customized videos, researchers are exploring various innovative approaches. In pursuit of model efficiency and generalization, zero-shot or tuning-free methods have emerged, exemplified by SUGAR (Zhou et al. 2024b), which leverages large-scale synthetic datasets. ConsisID (Yuan et al. 2025) offers a tuning-free method for identity preservation based on frequency decomposition. Still-Moving (Chefer et al. 2024) uniquely facilitates the customization of video models using solely image data through the training of spatial adapters. Another key research area is fine-grained control over motion and concepts. For instance, MotionBooth (Wu et al. 2024a) focuses on precise object and camera movement control. CustomTTT (Bi et al. 2025) employs test-time training to effectively combine multiple individually trained concepts (*e.g.*, appearance and motion LoRAs), while CustomCrafter (Wu et al. 2025b) aims to preserve the model’s intrinsic motion generation and concept composition capabilities during subject learning.

5.2.3 Summary

In this area, ID serves as a pivotal modality, utilizing conditional encoders and general control models (*e.g.*, Adapters) to achieve the core goal of personalized customization. While UNet remains the dominant backbone (*e.g.*, PersonalVideo (Li et al. 2025a)), there is a strategic shift toward DiT (*e.g.*, Movie Weaver (Liang et al. 2025)) to leverage global attention for improved temporal consistency.

Driven by the user demand for efficiency, the technical roadmap is transitioning from slow fine-tuning to Tuning-free approaches (*e.g.*, ConsisID (Yuan et al. 2025)), which directly aligns with the users’ demand for real-time. However, current limitations in multi-subject generation reveal a gap in the unified system of creative supports, underscoring the future need for advanced disentanglement and composition to resolve conflicts between different controllable methods (*e.g.*, conditional encoders and motion modules) without performance degradation.

5.3 Image Control

The image-guided video generation aims to generate a video from a given reference image. This form of control is crucial for producing videos that are not only visually consistent with the reference but also maintain coherence across temporal frames. By conditioning the generation process on reference image features, models can generate video content with enhanced visual alignment, stylistic fidelity, and semantic relevance to the source image. Approaches such as TRIP (Zhang et al. 2024f), OmniTokenizer (Wang et al. 2024d), Emu Video (Girdhar et al. 2023), and I2V-Adapter (Guo et al. 2024b) focus on addressing the insufficient consistency between the given image and the generated video, constructing a unified, high-quality video generation framework. To further construct a comprehensive video generation model, STIV (Lin et al. 2025c) and Hunyuan Video (Kong et al. 2024) are based on the Diffusion Transformer (DiT) framework with the image condition, which perform better in both T2V and I2V tasks.

Another challenge is to keep the alignment between the reference and the given prompt. Specifically, at the start of the generated video, the image always plays a more crucial role for the details, style, and the object location; however, in the later stage, the effects of the reference image can easily be weakened by the prompt. To solve this problem, DreamVideo (Wang et al. 2025b) introduces an Image Retention module, maintaining both information from the input image and prompt. AID (Xing et al. 2025b) incorporates an MLLM (Multimodal Large Language Model) and DQFormer (Dual Query Transformer) to predict future frames based on the given key frame. DynamiCrafter (Xing et al. 2024a) utilizes a dual-stream image injection paradigm to adapt various applications (*e.g.*, animation, storytelling). Techniques like Cond image leak (Zhao et al. 2024b) and EDG (Tian et al. 2025) aim to combat the issue of limited motion degrees and the unexpected motion with the prompt in the generated video. However, for the multi-object scenarios and the long-term video generation, the accuracy and consistency of the motion are also a challenge. Research like Through-The-Mask (Yariv et al. 2025), Lumiere (Bar-Tal et al. 2024), TI2V-Zero (Ni et al. 2024), and EasyAnimate (Xu et al. 2024b) attempt to utilize the powerful basic model (*e.g.*, DiT) or improve the computation paradigm (*e.g.*, compute all frames once and employ a special inversion strategy) while achieving a superior inference efficiency.

Despite those remarkable achievements, image condition can also be easily combined with various conditions such as BBox, motion, and depth. Motion-I2V (Shi et al. 2024b) first deduces the potential motion from the reference image, and then uses both predicted motion and image to generate high-quality video. Decoupled I2V (Shen et al. 2024) disentangles the motion vector and the image to obtain a memory-efficient and consistent video generation method. Similarly, Motion-Stone (Shi et al. 2025a) also models the motion by the motion head and injects it into the Diffusion Transformer. Besides, MoVideo (Liang et al. 2024) utilizes the motion and keyframe with the depth and optical flow for video generation. ST-I2V (Zhuo et al. 2023) cropped several areas from the reference image, using them to maintain the semantic information of the video.

Differently, PhysGen (Liu et al. 2024d) considers the physical properties like mass or elasticity, and external factors such as forces and environmental conditions. By giving input force and images as conditions, it can generate a video without a training process. HoloTime (Zhou et al. 2025) introduces a 360-degree 4D scene generation framework to reconstruct high-quality 4D scene video. ConsistI2V (Ren et al. 2024a) concatenates the input image to the noise and optimizes the attention calculation, not only performing better in consistency, but also supports various condition inputs (*e.g.*, layout, camera). To achieve more user-friendly video generation, video doodles (Yu et al. 2023) and Follow-Your-Click (Ma et al. 2025e) allow users to insert hand-drawn animations and provide a click to select which area to move, benefiting from better interactivity. Additionally, some work has explored using an image as the reference for other model architectures or tasks. NOVA (Deng et al. 2024) reformulates the video generation task as a non-quantized autoregressive modeling of temporal and spatial prediction, achieving a superior performance with fewer parameters. VideoPanda (Xie et al. 2025) also introduces a long video generation framework using auto-regression. Structure and Content-Guided Video Synthesis (Esser et al. 2023) extends the image condition to the video editing and video frame interpolation applications, respectively.

In summary, image control utilizes conditional encoders to map visual features into the backbone, fulfilling the user need for personalization and customization through high-fidelity visual alignment. In terms of the foundation models, the framework highlights a dynamic architectural ecosystem. Traditional UNet structures (*e.g.*, I2V-Adapter (Guo et al. 2024b)) are increasingly being complemented by DiT (*e.g.*, STIV (Lin et al. 2025c)) and Autoregressive models (*e.g.*, NOVA (Deng et al. 2024)), enabling superior scalability and temporal coherence for complex generation tasks. However, persistent challenges in multi-object scenarios and prompt-image alignment suggest a continued need for better explainability to help users understand how models balance conflicting inputs during the generation process.

5.4 Temporal Control

Temporal control in video generation refers to the ability to regulate the evolution of motion and timing across frames, ensuring that the generated content follows specific temporal dynamics. This form of control is crucial for synthesizing videos with coherent motion patterns, realistic pacing, and causally consistent frame transitions. By conditioning the generation process on temporal signals such as trajectories, action sequences, flow, or camera movements, models can produce videos that align with user-defined temporal intent.

5.4.1 Flow-Guided Generation

Flow-guided control utilizes optical flow or motion field representations to guide video synthesis, ensuring realistic temporal consistency and smooth object dynamics across frames. By incorporating motion cues, these methods provide finer control over movement and transitions, making them essential for coherent video generation tasks like animation and dynamic scene modeling.

Flow guidance can be specified through optical flow fields, motion trajectories, or displacement representations, either explicitly computed or implicitly learned. For example, Motion-I2V (Shi et al. 2024b) improves visual consistency by conditioning on motion trajectories, while MOFA-Video (Niu et al. 2024) animates static images using user-defined motion fields.

Recent approaches integrate flow-based guidance into generative architectures. I2VControl (Feng et al. 2025c) ensures temporal coherence through disentangled flow representations, while MCDiff (Chen et al. 2023c) employs motion-conditioned diffusion models for precise motion control. MOFA-Video (Niu et al. 2024) aligns motion fields for controllable animation, and Motion-I2V (Shi et al. 2024b) emphasizes flow-based trajectory conditioning to enhance consistency.

Challenges include maintaining motion consistency in complex scenes, avoiding artifacts from inaccurate flow, and efficiently integrating flow information without computational overhead. Techniques like MCDiff (Chen et al. 2023c) and Motion-I2V (Shi et al. 2024b) address these issues but highlight the need for further advancements in flow-guided video generation.

5.4.2 Trajectory-Guided Generation

Trajectory-guided video generation allows users to specify the motion paths of one or more objects or image regions across video frames. Applications include directing character movements along a specific route, animating inanimate objects with desired kinematics, or ensuring specific interactions between entities based on their spatial paths.

Trajectory input can take various forms: a sequence of 2D or 3D coordinates $\{\mathbf{x}_t, \mathbf{y}_t, (\mathbf{z}_t)\}_{t=1}^T$ for a point of interest, bounding box tracks $\{\text{bbox}_t\}_{t=1}^T$ defining the extent and location of an object over time (as in Boximator (Wang et al. 2024b)), user-sketched paths in an initial frame, or even textual descriptions of motion (*e.g.*, “move object A from left to right”). Drag-based interfaces, as explored in DragAnything (Wu et al. 2024b) and DragEntity (Wan et al. 2024), allow interactive specification of start and end points for object parts.

Many recent methods leverage diffusion models for trajectory control. MotionBooth (Wu et al. 2024a) and Motion-I2V (Shi et al. 2024b) focus on generating video from an image based on motion prompts or trajectories. TrailBlazer (Ma et al. 2024b) and Direct-a-Video (Yang et al. 2024b) (which also handles camera control) enable explicit path following. Peekaboo (Jain et al. 2024) controls object appearance and disappearance along trajectories. MOFA-Video (Niu et al. 2024) focuses on fine-grained trajectory control. Trajectory information is often encoded as a sequence of coordinates or heatmaps and is used to guide the cross-attention layers or directly modify the latent representations in diffusion models. Warping fields or flow-based methods can also be guided by trajectories. MCDiff (Chen et al. 2023c) introduced motion-controlled diffusion. More advanced concepts like Perception-as-Control (Chen et al. 2025e) and Motion Canvas (Xing et al. 2025a) provide sophisticated frameworks. LeviTor (Wang et al. 2025d) and LayerAnimate (Yang et al. 2025c) explore layered representations for controllable animation along trajectories.

Key challenges include generating motion that is not only accurate to the specified trajectory but also appears natural and physically plausible, avoiding jerky or sliding movements. Ensuring that the guided object interacts realistically with other scene elements and the background

Table 5: Quantitative comparison of camera-guided video generation on 3D static and 4D dynamic scenes, using public and internet data. We evaluate generation quality (FID, CLIP_{sim} for static; FVD, CLIP-V_{sim} for dynamic) and trajectory accuracy (ATE, RPE-T, RPE-R). All methods use official code with identical inputs. \uparrow indicates higher is better, \downarrow indicates lower is better. The best and secondary results are in bold and underline, respectively.

Methods	Generation Quality				Trajectory Accuracy					
	Static		Dynamic		Static			Dynamic		
	FID \downarrow	CLIP _{sim} \uparrow	FVD \downarrow	CLIP-V _{sim} \uparrow	ATE \downarrow	RPE-T \downarrow	RPE-R \downarrow	ATE \downarrow	RPE-T \downarrow	RPE-R \downarrow
See3D (Ma et al. 2025a)	123.26	0.941	—	—	0.091	0.089	0.250	—	—	—
ViewCrafter (Yu et al. 2025a)	117.50	0.930	—	—	0.236	0.315	0.728	—	—	—
ViewExtrapolator (Liu et al. 2024b)	125.50	0.930	108.48	0.913	0.183	0.260	0.882	1.040	1.208	4.750
TrajectoryAttention (Xiao et al. 2025b)	122.37	0.920	106.94	0.911	0.159	0.238	0.532	0.605	1.238	3.560
TrajectoryCrafter (YU et al. 2025)	111.49	0.910	97.31	0.923	0.090	0.152	0.267	0.431	1.078	8.950
NVS-Solver (You et al. 2024)	118.64	0.937	—	—	0.224	0.268	1.056	—	—	—
WorldForge(Song et al. 2025a)	96.08	0.948	93.17	0.938	0.077	0.086	0.221	0.527	0.826	2.690

is difficult. Handling complex, intersecting, or occluded trajectories for multiple objects (as in Collaborative Video Diffusion (Kuang et al. 2024)) requires sophisticated reasoning. Maintaining long-range temporal consistency of the objects’ appearance and the trajectory adherence over many frames is also crucial. Tora (Zhang et al. 2025o) and MagicMotion (Li et al. 2025b) aim to address some of these complexities, with works like I2VControl (Feng et al. 2025c) focusing on image-to-video with trajectory control.

5.4.3 Camera-Guided Generation

Camera-guided video generation focuses on controlling the virtual camera’s parameters to dictate the viewpoint, motion (*e.g.*, pan, tilt, zoom, dolly, crane shots), and potentially intrinsic parameters (*e.g.*, focal length) during video synthesis. This is essential for achieving specific cinematic styles, narrative perspectives, and dynamic visual effects, offering filmmakers and content creators fine-grained control over how a scene is presented.

Camera control signals can be provided as explicit sequences of camera poses (extrinsics: rotation \mathbf{R}_t and translation \mathbf{T}_t , and intrinsics \mathbf{K}_t if they vary), relative transformations between frames, or higher-level textual descriptions of camera movements (*e.g.*, “dolly zoom out from the character”). Motion Prompting (Geng et al. 2025) explores textual descriptions for camera and object motion. Recent advancements heavily leverage diffusion models conditioned on camera parameters. CameraCtrl (He et al. 2024a) and its successor CameraCtrl II (He et al. 2025) provide

explicit control over camera trajectories. Direct-a-Video (Yang et al. 2024b), also mentioned for trajectory control, supports camera path guidance. ViewCrafter (Yu et al. 2025a) focuses on generating novel views based on camera inputs. CamCo (Xu et al. 2024a) and Trajectory Attention (Xiao et al. 2025b) explore camera control in specific contexts. Uni3C (Cao et al. 2025) explores the joint control of human motion and camera trajectory. The camera parameters are typically used to define projection matrices that transform latent scene representations or guide the sampling of features to render the scene from the specified viewpoint. EgoSim (Yu et al. 2025b) focuses on egocentric camera simulation. Approaches like I2VControl (Feng et al. 2025c), CamI2V (Zheng et al. 2024), and RealCam-I2V (Li et al. 2025d) specialize in image-to-video generation with camera control. More advanced systems like 3DTraj-Master (Fu et al. 2024), MotionFlow (list not fully provided 2025), and CineMaster (Wang et al. 2025f) aim for comprehensive 3D-aware camera and motion control.

Challenges include maintaining 3D scene consistency and avoiding disocclusion artifacts or distorted geometry when the camera undergoes significant movement or viewpoint changes. Generating high-fidelity novel views that are consistent with previous frames is difficult. Ensuring smooth and natural camera transitions, free from jitter or abrupt changes, is crucial for cinematic quality. Providing intuitive user interfaces for specifying complex 3D camera paths and parameters remains an open area. ReCamMaster (Bai et al. 2025) aims to improve the quality and control of camera movements. Techniques like VD3D (Bahmani

et al. 2025b), Aether (Zhu et al. 2025a), and GenDoP (Zhang et al. 2025k) explore complex 3D scene understanding and camera path generation. The latest work Voyager (Huang et al. 2025b) introduces a world caching scheme and smooth video sampling, and Follow-Your-Creation (Ma et al. 2025d) utilizes the temporal pack inference strategy to keep the long-term consistency, respectively.

5.4.4 Motion-Guided Generation

Motion-guided video generation is the task of generating videos where a predefined motion concept—such as walking, dancing, waving, or jumping—is explicitly provided as input and used to guide the synthesized motion in the output video. This motion concept represents a coherent and semantically meaningful behavior, typically extracted from one or more reference video clips or motion representations. Unlike standard text-to-video generation that freely imagines motion from prompts, motion-guided methods aim to transfer concrete motion trajectories or dynamics—often in the form of pose sequences, optical flow, or velocity fields—into newly generated visual content. This enables fine-grained control of subject behavior in applications such as character animation, avatar customization, and creative video editing.

The work generally utilizes an explicit motion signal as the control anchor, with an optional text prompt defining scene semantics. Motion control signals vary in format: structured pose sequences like MoTrans (Li et al. 2024b) and MotionPrompting (Geng et al. 2025), hand mask sequences like InterDyn (Akkerman et al. 2025), or optical flow fields like OnlyFlow (Koroglu et al. 2025) provide dense or sparse motion descriptors. Some works directly use reference videos as motion priors, including DreamMotion (Jeong et al. 2025a), DreamRelation (Wei et al. 2025), Customize-A-Video (Ren et al. 2024b), Dual-Real (Wang et al. 2025g), and VideoMage (Huang et al. 2025a). Other works encode abstract fields, such as velocity maps (Liao et al. 2025) or relational cues. Even in motion-guided settings, text prompts are retained to define the appearance,

layout, or objects involved in the scene. Furthermore, a growing subset of works supports zero-shot or training-free transfer, including MotionClone (Ling et al. 2025) and VMC (Jeong et al. 2024), by leveraging alignment in latent spaces or score distillation techniques.

To realize controllable generation, a range of architectural designs and mechanisms have been introduced, often combining motion-aware modules with large-scale diffusion backbones. DreamActor-M1 (Luo et al. 2025b) injects motion tokens into DiT transformers for coherent spatiotemporal control. InterDyn (Akkerman et al. 2025) performs targeted gesture synthesis via attention-guided injection of hand masks. DreamMotion (Jeong et al. 2025a) distills score maps from reference videos to enforce temporal behavior. DreamRelation (Wei et al. 2025) models trajectory consistency using relation fields between appearance regions. Methods like MotionAgent (Liao et al. 2025) and MotionPrompting (Geng et al. 2025) formulate motion as fields or anchor prompts for fine control, while SMA (Park et al. 2025) and Diffusion as Shader (Gu et al. 2025) propose spectral and 3D-aware conditioning mechanisms. Meanwhile, MotionMatcher (Wu et al. 2025d), Motionbooth (Wu et al. 2024a), and MotionDirector (Zhao et al. 2025c) align motion tokens in latent space, and Customize-A-Video (Ren et al. 2024b) learns a one-shot controller from minimal supervision.

Despite these advances, motion-guided generation still faces significant challenges. Motion signals can be noisy, ambiguous, or misaligned with the model’s internal representation, often causing motion artifacts or flicker. Recent work Motion-Pro (Zhang et al. 2025p) proposes a region-wise trajectory and motion mask to alleviate misinterpretation and achieve better motion control.

Furthermore, generalization to unseen motions or subjects remains limited, particularly in one-shot or zero-shot settings (*e.g.*, MotionClone (Ling et al. 2025), VMC (Jeong et al. 2024)). Evaluation also lags behind: although SST-EM (Biyala et al. 2025) introduces metrics for semantic, spatial, and temporal fidelity, widely accepted benchmarks and interpretability tools are under development.

Table 6: Quantitative comparisons of camera-guided video generation on T2V setting across RealEstate10K and DL3DV. Frame Consistency and Motion Strength are denoted as FC and MS. In addition. The rotation and translation error are denoted as RE and TE.

Method	RealEstate10K						DL3DV					
	FVD ↓	CLIPSIM ↑	FC ↑	MS ↑	RE ↓	TE ↓	FVD ↓	CLIPSIM ↑	FC ↑	MS ↑	RE ↓	TE ↓
MotionCtrl (Wang et al. 2024h)	506.9	0.2744	0.9471	0.9734	2.90	1.06	746.7	0.2033	0.9448	0.9462	1.17	3.23
CameraCtrl (He et al. 2024a)	426.8	0.2734	0.9526	0.9750	2.68	0.98	675.5	0.2027	0.9518	0.9610	1.01	4.00
AC3D (Bahmani et al. 2025a)	415.6	0.3044	0.9496	0.9811	2.67	0.38	452.8	0.2078	0.9529	0.9869	1.06	2.11
DualCamCtrl (Zhang et al. 2025a)	408.1	0.3154	0.9540	0.9918	1.23	0.25	427.4	0.2090	0.9544	0.9807	0.83	1.53

5.4.5 Summary

Temporal Control activates the dynamic layer of the framework, employing motion modules (*e.g.*, Trajectory Samplers) and camera control modules to translate static concepts into coherent narratives. This modality is central to satisfying the user needs for automation and augmentation, as it replaces manual keyframing with semantic guidance like flow fields or drag-based interactions (*e.g.*, DragAnything (Wu et al. 2024b)), thereby streamlining the creation of complex scenes.

Regarding the architecture, while UNet models (*e.g.*, Motion-I2V (Shi et al. 2024b)) excel at short-term pixel-level flow dynamics, the complexity of Video-to-4D generation is necessitating a move towards DiT (*e.g.*, DreamActor-M1 (Luo et al. 2025b)) and AR backbones. These architectures are favored not just for size, but for their ability to model long-term causal dependencies and 3D geometric consistency (*e.g.*, Voyager (Huang et al. 2025b)), which are critical when coordinating sophisticated camera paths with object interactions over extended durations.

In the future, the trend of temporal control is moving towards integrated control where temporal signals (Trajectories/Flow) need to coexist with spatial controls (Structure/ID). Current general control models often struggle to merge these distinct signals seamlessly, highlighting a user demand for better interactive editing tools (*e.g.*, drag-based interfaces like DragAnything (Wu et al. 2024b)) that provide real-time preview of motion paths before generation. Future frameworks will likely prioritize universal control mechanisms to unify these diverse motion inputs into a single interface.

5.5 Audio Control

Audio control in video generation refers to the ability to synthesize videos from audio signals such as speech, music, or general sound. This form of control enables models to generate temporally aligned and semantically coherent videos based on audio cues. A subdomain is voice control, which focuses on generating talking face videos or personalized portrait animations conditioned on speech input. Another direction is sound control, which utilizes broader forms of audio (*e.g.*, music or ambient sound) to guide the generation with more general content beyond facial animation.

5.5.1 Voice-Guided Generation

Voice-guided video generation aims to use both audio and image as conditions to create realistic videos, especially for the talking portrait video generation tasks, which is considered one of the most crucial trends in the applications of the human-computer interaction field. The input embeddings of the sound condition sequences S_t firstly are extracted commonly by the pretrained encoder like wav2vec (Schneider et al. 2019). In order to acquire richer semantic information from the input audio, several works (Xu et al. 2024c; Wang et al. 2024a; Tian et al. 2024b) have concatenated features from different encoder layers for the f -th frame S^f and injected them into the diffusion processes by the cross-attention layers.

The core issue in talking portrait video generation tasks is capturing and estimating the actual face expressions from the various voice inputs, such as the movements of lips, head poses, and eyebrows, which influence the quality of the video significantly. Recently, (Sun et al. 2025), (Wei et al. 2024a) propose a two-stage architecture based on 3D mesh as the intermediate representation. EDTalk (Tan et al. 2024a)

Table 7: Quantitative comparison with state-of-the-art methods averaging on ambient sound-video, speech-video, and complex scene-video generation.↑ indicates higher is better, ↓ indicates lower is better.

Method	Video Quality & Coherence					Audio Fidelity & Quality						Audio-Visual Synchronization			
	AQ ↑	IQ ↑	DD ↑	MS ↑	ID ↑	PQ ↑	PC ↓	CE ↑	CU ↑	WER ↓	IB-A ↑	Sync-C ↑	Sync-D ↓	DeSync ↓	IB ↑
MM-Diffusion (Ruan et al. 2023)	0.32	0.43	0.13	0.99	-	5.37	4.07	4.27	5.89	-	-	-	-	-	0.12
JavisDiT (Liu et al. 2025c)	0.34	0.53	0.38	0.99	0.38	5.46	2.24	3.19	4.54	1.00	0.14	0.89	11.62	1.13	0.18
UniVerse-1 (Wang et al. 2025c)	0.52	0.67	0.24	0.99	0.89	5.52	<u>2.13</u>	3.63	4.84	<u>0.24</u>	0.07	0.97	10.71	<u>1.10</u>	0.12
Ovi (Low et al. 2025)	0.57	<u>0.65</u>	0.34	0.99	<u>0.90</u>	<u>6.19</u>	<u>2.13</u>	<u>4.44</u>	<u>5.84</u>	0.49	<u>0.12</u>	<u>4.04</u>	<u>9.62</u>	1.14	<u>0.18</u>
Harmony (Hu et al. 2025b)	0.59	<u>0.65</u>	<u>0.36</u>	0.99	0.91	6.39	2.05	4.73	5.67	<u>0.15</u>	<u>0.12</u>	5.61	7.53	0.92	0.19

has disentangled the motion space into three distinct latent spaces and designs an audio-to-motion module to predict different expressions. EMO (Tian et al. 2024b) designs the audio feature extractor, face locator, and speed layers, combining with the ReferenceNet (Hu 2024) and motion frame module, which ensures the performance of generated videos. EchoMimic (Chen et al. 2025f) and EchoMimicV2 (Meng et al. 2025a) also use the ReferenceNet to extract the features from the input image, generating a better portrait video.

Different from other conditions like images, layouts, and sketches, the strength of sound is weaker, hindering the control effectiveness during the inference process. Therefore, addressing the conflict between various conditions is another challenge for voice-guided video generation. V-Express (Wang et al. 2024a) proposes a progressive training strategy and conditional dropout operations, which are beneficial for those weak control conditions. MOFA-Video (Niu et al. 2024) designs a MOFA-Adapter to animate a reference image into a video under various control conditions. To balance the control strength between different conditions, MegActor- Σ (Yang et al. 2025a) proposes a mixed-modal DiT, ACTalker (Hong et al. 2025) introduced a mask-drop strategy, while MotionCraft (Bian et al. 2025b) and OmniHuman-1 (Lin et al. 2025b) optimize the training strategy of the multimodal generation framework.

Besides those challenges, there are still some issues, such as limitations in style variation generation on efficiency. For the former one, SVP (Tan et al. 2024b) models the intrinsic style of the source video to alleviate this problem. For the latter one, (Ki et al. 2024; Xu et al. 2024d; Lin et al. 2025a) improve the real-time performance for the video-chat application. Finally,

as for the consistency in long-term video generation, MCDM (Shen et al. 2025) designs two clip-motion-prior modules and a memory-efficient attention mechanism. In contrast, TexTalker (Li et al. 2025e) focuses on the dynamic texture for the high-fidelity talking heads generation.

5.5.2 Sound-Guided Generation

Sound-guided video generation involves creating videos where visual elements are synchronized with general sound inputs like music, speech, or ambient noises. This method leverages temporal and emotional properties of the sound to guide the visual content, creating videos that reflect not just the overall semantics of the audio but also its dynamic features. In this approach, sound serves as a critical modality for determining the motion, scene structure, and pace of the generated video.

In sound-guided video generation, sound inputs range from specific sounds like animal calls and music to general audio cues. Various audio feature extraction methods, such as Mel frequency cepstral coefficients (MFCCs), Mel spectrograms, BEATs, and CLAP, are used to capture texture, pitch, rhythm, and emotional tones. For instance, AV-Link (Haji-Ali et al. 2025) and ASVA (Zhang et al. 2024b) synchronize visuals with audio by extracting rhythmic, spectral, and emotional properties using BEATs, CLAP, and MFCCs or Mel-Spectrograms. TA2V (Zhao et al. 2024a) further integrates text, using BEATs to guide video generation, showcasing how audio dynamics shape visual output.

Recent advancements in sound-guided video generation, such as AV-Link (Haji-Ali et al. 2025) and ASVA (Zhang et al. 2024b), have shifted toward diffusion models for more accurate synchronization, using temporally-aligned activations and audio features to ensure better alignment.

TA2V (Zhao et al. 2024a) further enhances this by incorporating both text and audio inputs, conditioning video motion and emotion. Motion-Craft (Bian et al. 2025b) introduces a dual-branch architecture for rhythmic and semantic alignment, while DAA2V (Yariv et al. 2024) refines the process with text-to-video models for improved temporal and semantic alignment.

Challenges in sound-guided video generation include maintaining synchronization between audio and video, especially with diverse sound types. Solutions like AV-Link(Haji-Ali et al. 2025) and ASVA(Zhang et al. 2024b) use temporal attention mechanisms to address this. TA2V(Zhao et al. 2024a) tackles the complexity of coordinating both text and audio, while DAA2V(Yariv et al. 2024) provides insights on maintaining synchronization across different audio types.

5.5.3 Summary

In this domain, audio functions as a temporal-dynamic anchor, guiding the rhythmic progression of visual content. A primary challenge is fine-grained temporal alignment to prevent synchronization drift. This issue has driven a shift from UNet-based pipelines (*e.g.*, *EchoMimic* (Chen et al. 2025f)) to unified joint generation. Spearheading this evolution, recent works like *MegActor- Σ* (Yang et al. 2025a) leverage Diffusion Transformers (DiT) to ensure superior long-range coherence.

Driven by users' demand for high-fidelity interaction, current research tackles a critical control imbalance, where dominant visual features frequently suppress subtle audio cues. It necessitates balanced training strategies, such as the conditional dropout in *V-Express* (Wang et al. 2024a). However, the separation of structural voice and semantic sound control remains a limitation, suggesting a future move toward holistic audio-visual physics to unify micro and macro dynamics.

5.6 Other Control

Other control in video generation refers to a set of control mechanisms beyond identity, structure, and motion, enabling more diverse and fine-grained alignment between user intent and generated content. This includes text rendering, which guides generation to produce videos aligned with specific textual elements; style control, which

transfers visual characteristics from a reference image or artistic domain to the synthesized video; point control, which enables spatial manipulation through user-specified keypoints or dragging operations; and BEV control, which is especially relevant in autonomous driving scenarios where bird's-eye view maps or semantic layouts are used to guide video synthesis.

5.6.1 Text Rendering

Text rendering in the video domain targets synthesizing vivid video with aligned text, which has widespread applications in various forms, including advertisements and movies. Previous works (Jiang et al. 2024b; Shi et al. 2024a, 2025b; Zeng et al. 2024) leverage the CLIP (Radford et al. 2021) text encoder to encode prompts and use 2D VAE (Jiang et al. 2025b) to compress the video from RGB to latent space. They fail to generate clear and aligned text. Text-Animator (Liu et al. 2024c) designs a text embedding injection module, a camera control module, and a text refinement module to improve both visual text structures and stability. Recently, Wan (Wan et al. 2025) has proposed a series of powerful foundation models, including a 1.3B model and a 14B model, which expand the application to various areas such as T2V, text rendering, audio generation, and numerous other downstream tasks.

5.6.2 Style-Guided Generation

Style-guided video generation transfers an artistic style from a reference image or video to a target video, which is central for creative expression and for achieving specific visual aesthetics. The field has progressed from early per-frame stylization with post-hoc temporal smoothing (Chen et al. 2017; Huang et al. 2017) to efficient and flexible frameworks for arbitrary style transfer (Song et al. 2025e; Guo et al. 2025; Song et al. 2024c,b; Zhang et al. 2025n; Chen et al. 2025d; Zhang et al. 2024e; Deng et al. 2021; Gao et al. 2020). With the development of generative models, GAN-based approaches support specialized applications such as video toonification (PYang et al. 2022) and video translation (Yang et al. 2023b), while more recent diffusion models, including Rerender A Video (Yang et al. 2023a) and StyleCrafter (Liu et al. 2024a), now form the leading approach for high-fidelity video stylization. Building on this,

FRESCO (Yang et al. 2024a) and StyleMaster (Ye et al. 2025) introduce more fine-grained stylization controls, while UniVST (Song et al. 2024a) aims at a general framework for diverse artistic styles. Remaining challenges include temporal consistency, preservation of content details under strong stylization, and intuitive user controls.

5.6.3 Point-Guided Generation

Point-guided video generation provides fine-grained control by using sparse user-defined point trajectories to animate specific objects or regions, and the main challenge is to extend these constraints in a stable way over space and time. GAN-based methods such as Point-to-Point Video Generation (Wang et al. 2019) and ImaGINator (Wang et al. 2020) first explore sparse point inputs for controllable video synthesis, while later work often conditions pre-trained diffusion models on such inputs. Track4Gen (Jeong et al. 2025b) focuses on robustly tracking user-specified points so that the generated content follows the desired trajectories, and related frameworks such as Diffusion as Shader (Gu et al. 2025) and GS-DiT (Bian et al. 2025a) show that sparse user input can steer more complex scene properties. These methods typically encode points as heatmaps or use them to guide attention mechanisms, translating sparse input into coherent and dynamic outputs. Key challenges include physically plausible deformations, long-term temporal consistency from minimal input, and handling complex object interactions and occlusions.

5.6.4 BEV-Guided Generation

Bird’s-Eye View (BEV) guided video generation targets controllable and realistic driving scenarios for autonomous systems, where a sequence of top-down semantic maps $B = \{b_1, b_2, \dots, b_T\}$ specifies scene layout and object dynamics. Early GAN-based methods such as DriveGAN (Kim et al. 2021) study BEV-conditioned video synthesis for realistic driving sequences, and recent work establishes diffusion models as a central paradigm for translating abstract BEV representations into photorealistic videos, as illustrated by MagicDrive (Gao et al. 2023), MagicDrive-V2 (Gao et al. 2025), DriveDreamer-2 (Zhao et al. 2025a), DiVE (Jiang et al. 2024a), Delphi (Ma et al. 2024a), DreamForge (Mei et al. 2024),

Panacea (Wen et al. 2024), and Seeing Beyond Views (Lu et al. 2024). Despite this progress, key challenges remain in geometric consistency between BEV inputs and rendered videos, realistic appearance and lighting, accurate modeling of complex dynamic interactions, and the semantic gap between abstract BEV representations and fine-grained visual details.

5.6.5 Summary

Other control modalities, including text rendering, style transfer, point tracking, and BEV conditioning, provide fine-grained and semantically aligned control over video generation beyond identity, structure, and motion. In our framework, these modalities connect foundation models with high-level creative tools through conditional encoders and control modules such as adapters and cross-attention mechanisms, which inject sparse or abstract guidance into the generative process. The design trend is shifting from mainly UNet-based architectures (for example, ControlNet-based stylization) toward DiT-based backbones, as in Wan and advanced BEV models, to better capture global semantic context and complex spatio-temporal dependencies for tasks such as coherent text spelling and autonomous driving simulation. From a user perspective, these controls support personalization through style and text and real-time preview and adjustment through point dragging, while turning these mechanisms into intuitive, transparent tools with effective automatic suggestions remains an important direction for future research.

5.7 Universal Control

Universal Control refers to the ability to flexibly manage and integrate various types of input conditions—such as text, spatial features, and temporal information—to generate videos that meet specific user requirements. This method allows for high customization, enabling users to influence multiple aspects of the generated video simultaneously, such as object behavior, motion, camera angles, and scene layout. By unifying different conditions into a single, coherent framework, universal control facilitates comprehensive video synthesis with enhanced flexibility and precision.

In universal-guided video generation, the model typically incorporates a range of inputs,

including text, spatial conditions (such as camera poses, depth maps, and human poses), and temporal signals (including motion sequences and video clips). For example, frameworks like VideoComposer (Wang et al. 2023a) utilize motion vectors or sketches as explicit control signals, guiding the video generation process. Similarly, Any2Caption (Wu et al. 2025a) employs multimodal inputs, such as text, camera poses, and human motions, which are interpreted into structured captions to drive video synthesis. FullDiT (Ju et al. 2025) integrates these various conditions using a unified attention mechanism, processing text, camera views, depth information, and identity conditions simultaneously, thus enabling seamless video generation.

Several recent frameworks highlight different ways to achieve this. For instance, VideoComposer (Wang et al. 2023a) relies on a dedicated spatio-temporal condition encoder to blend motion cues, depth information, and text prompts inside a latent diffusion pipeline. Any2Caption (Wu et al. 2025a), on the other hand, takes a more indirect route: it first translates diverse inputs into detailed, structured textual captions using a multimodal LLM, then feeds those captions to a standard video generator. FullDiT (Ju et al. 2025) pushes direct integration further by employing full spatiotemporal attention within a Diffusion Transformer architecture, letting it handle text, camera poses, depth, and even identity constraints all at once. Notably, in the autoregressive camp, InfinityStar (Liu et al. 2025b) stands out—it simply tokenizes everything (visual content and controls alike) into a unified discrete sequence, enabling native multi-condition control for tasks ranging from text-to-video to image animation and long-sequence extension, all without extra adapters.

Despite the progress made in universal-guided control, several challenges persist. One of the primary difficulties lies in balancing the integration of various conditions. Ensuring that each condition (*e.g.*, motion, text, camera) is harmoniously combined without one overshadowing the others remains a complex task. In addition, maintaining temporal consistency across frames presents another challenge, particularly when dealing with dynamic and varied motion patterns. Furthermore, evaluation continues to be an unresolved issue—while benchmarks like FullBench for

FullDiT have been introduced, the development of comprehensive and standardized evaluation criteria for universal control across all modalities is still underway.

In summary, universal control uses advanced conditioning—from encoders and attention fusion to token unification—to combine signals like text, poses, depth, and motion. This lets users customize content, style, motion, and layout all at once, far surpassing what single-condition models can do.

The architecture has shifted from bulky UNet-with-adapters designs to more scalable backbones, driven by the need for cleaner multi-condition handling and better scaling. Diffusion Transformers (*e.g.*, FullDiT (Ju et al. 2025)) and autoregressive models (*e.g.*, InfinityStar (Liu et al. 2025b)) now lead, offering natural fusion of controls, support for long/high-res sequences, and either high fidelity or fast in-context flexibility. Caption approaches like Any2Caption (Wu et al. 2025a) and early composable encoders from VideoComposer (Wang et al. 2023a) remain useful, but trends favor integrated, interactive systems. Challenges in signal conflicts, long-term coherence, and interpretability persist, suggesting future progress will focus on hybrid designs and clearer user insight into control interactions.

6 Application

With the rapid advancement of video generative models, video generation techniques are increasingly being applied to real-world scenarios. As illustrated in Fig. 7, T2V models have demonstrated significant potential in various areas.

Video Completion and Inpainting. Video completion and inpainting aim to fill in missing areas in videos or remove unnecessary objects while maintaining spatial and temporal consistency. The techniques can be widely used in film post-production, historical video restoration, and everyday mobile video editing apps. Traditional methods often struggle with large occlusions or complex motion. In recent years, video generative models have shown strong generative capabilities in this field. DiffuEraser (Li et al. 2025g) is designed to fill occluded areas, combining prior information for initialization and weak conditional constraints to help reduce noise artifacts and suppress hallucination. AVID (Zhang et al. 2024g)



Fig. 7: Applications of Controllable Video Generation. The sample images are from DiffuEraser (Li et al. 2025g), VideoAnydoor (Tu et al. 2025), 4D-fy (Bahmani et al. 2024), Vista (Gao et al. 2024), spmem (Wu et al. 2025c), RoboMaster (Fu et al. 2025).

introduces motion modules and adjustable structural guidance, which support different types of repair under various levels of structural fidelity, capable of handling videos of any length, ensuring temporal consistency within the editing area.

Video Composition. Video composition aims to seamlessly embed target objects into existing videos while ensuring both temporal and spatial consistency. It is essential in advertising production, AR/VR environments, and creator tools such as short-video platforms. It allows users to insert virtual objects into real scenes, enabling applications like virtual product placement, AR furniture try-on, or creative storytelling with animated characters. It significantly reduces the effort needed to produce high-quality mixed-reality content. This task requires the inserted object to be well integrated with the background while preserving the details of the appearance of object. Recent advances in text-to-video diffusion models offer powerful control capabilities for this purpose. For example, MVOC (Wang et al. 2024f) reverses the corresponding noise features by DDIM inversion on each video object, then synthesizes and edits them to generate the first frame of the synthetic video. VideoAnydoor (Tu et al. 2025) is an end-to-end video object insertion framework that achieves high-fidelity detail retention and precise motion control through the ID extractor, Pixel Warper, and optimized training strategies.

Video-to-4D Generation. 4D generation is inherently more challenging than static 3D reconstruction, as it requires modeling both spatial geometry and temporal dynamics. Video generative models naturally provide strong motion priors by capturing scene dynamics over time, making them an emerging core technology for constructing high-quality 4D representations. It enables game studios, film creators, and cultural heritage organizations to build 4D digital assets from simple handheld videos, avoiding expensive multi-camera capture systems. It can also support robotic perception by providing realistic, dynamic 3D environments for simulation and training. 4Dfy (Bahmani et al. 2024) proposes a mixed score distillation method that combines multiple pre-trained diffusion models to generate 4D scenes with realistic appearance, structure, and motion. 4Real (Yu et al. 2024) utilizes a video diffusion model to generate reference videos and “freeze-time” videos, and learns a standard 3D representation based on deformable 3D Gaussian splatting and time-varying deformations.

Autonomous Vehicle and Generative World Model. World model is important in autonomous driving by simulating environments and guiding decision-making. Video generation models, with their ability to produce realistic visual data, are now increasingly used to improve perception and planning in autonomous systems. GAIA-2 (Russell et al. 2025) is a multi-view generative world model capable of producing high-resolution,

temporally and spatially consistent multi-camera videos from structured inputs such as vehicle dynamics, environmental conditions, and road semantics. DriVerse (Li et al. 2025f) generates high-fidelity driving simulation videos by using a single image and future trajectories, introduces multi-modal trajectory prompts to encode trajectories as text and spatial motion priors, and uses latent motion alignment to enhance the temporal consistency of dynamic objects. Drive-WM (Wang et al. 2023c) promotes spatial-temporal joint modeling through view factorization, enabling the generation of high-fidelity multi-view videos in autonomous driving scenarios. The powerful generative capability demonstrates the potential of applying world models to safe driving planning.

Embodied Artificial Intelligence. Video generative models can simulate future scenarios, enabling embodied agents to understand, predict, and plan interactive behaviors by generating visual information that aligns with semantic and physical constraints. This enhances the agents’ capabilities within the perception–decision–action loop. Gen2Act (Bharadhwaj et al. 2024) introduces a framework that uses video generation model to generate human operation videos. These videos are translated into executable robot actions via a policy model, addressing the challenge of generalizing to unseen objects and novel tasks. AVDC (Ko et al. 2023) leverages video diffusion model to synthesize hallucinated videos of robots performing actions. It extracts dense correspondences via optical flow within the synthesized video and combines them with initial depth estimates, thereby inferring robot actions. RoboMaster (Fu et al. 2025) generates robot manipulation videos through collaborative trajectory control. It addresses the feature entanglement problem in multi-object interaction by decomposing the process into three stages and integrating object embeddings with appearance and shape perception. This&That (Wang et al. 2025a) enhances robot planning and execution by generating task videos conditioned on both language and gestures. Compared to language-only methods, this multimodal approach enables more accurate interpretation of user intentions.

7 Discussion and Future Work

Controllable video generation has advanced rapidly (Tab. 2), yet building robust and scalable systems that support diverse controls remains challenging. Current methods often optimize a subset of constraints (*e.g.*, structure, temporal consistency, style), leading to unavoidable trade-offs when multiple conditions must be satisfied simultaneously. We highlight three promising directions for future research.

7.1 Unified Multi-Condition Control and Constraint Orchestration

A key bottleneck is compositional controllability: real applications usually require jointly satisfying temporal coherence, spatial/structural fidelity, motion dynamics, and stylistic consistency. Existing works typically emphasize one aspect at a time—*e.g.*, structure-driven generation such as Follow-Your-Pose (Ma et al. 2024c) and DriveDreamer (Zhao et al. 2025a), temporal control such as MotionBooth (Wu et al. 2024a) and Direct-A-Video (Yang et al. 2024b), and image/style-guided generation such as VideoCrafter1 (Chen et al. 2023b) and Lumiere (Bar-Tal et al. 2024). As a result, enforcing one constraint often degrades others.

Future systems should move toward constraint orchestration: (i) hierarchical or modular control frameworks that compose heterogeneous conditions and explicitly manage conflicts; (ii) adaptive control policies that can re-weight constraints based on user intent and content (*e.g.*, prioritizing temporal smoothness for action sequences while emphasizing style for artistic videos). VideoComposer (Wang et al. 2023a) is an early step toward universal conditioning, but stronger mechanisms are needed to reliably handle complex, mixed user inputs. Compositional diffusion and unified architectures (*e.g.*, FullDiT (Ju et al. 2025)) further suggest the feasibility of multi-condition integration, yet open problems remain in balancing competing constraints and maintaining computational efficiency.

7.2 Unified Video Reasoning and Generation

Another direction is to couple reasoning with generation (Chen et al. 2025b,c,a). LLMs can serve as a semantic control layer that translates high-level intent into concrete, low-level conditions and supports iterative refinement through multi-turn interaction. Recent systems such as Phantom (Liu et al. 2025d) demonstrate progress toward semantically rich, prompt-driven generation, but robust alignment between reasoning modules and video generators is still limited, especially under complex instructions and long-horizon consistency requirements.

A promising route is multi-modal alignment that allows users to combine text with other signals while keeping the interaction natural and editable. In this context, Multi-Modal LLM (MLLM) agents can act as a unified interface for planning, grounding, and refinement: they can parse intent, select or synthesize appropriate controls, and iteratively update constraints based on user feedback. Key challenges include (i) reliable grounding and verification to prevent instruction drift and hallucinated content; (ii) bias and safety mitigation in open-ended generation; and (iii) efficiency, *e.g.*, lightweight adapters or cross-modal modules that reduce compute overhead while preserving controllability.

7.3 Hybrid and Scalable Autoregressive Methods

Scaling to long videos with strong temporal coherence remains difficult. Diffusion models produce high-quality frames but may suffer from accumulated inconsistencies over long durations, whereas autoregressive generation explicitly models temporal dependencies and can improve coherence (*e.g.*, NOVA (Deng et al. 2024)), at the cost of higher latency and compute.

Hybrid designs that combine diffusion and autoregressive components are therefore attractive: diffusion can generate globally consistent latent representations, while lightweight autoregressive modules enforce sequential continuity. Multi-scale modeling is also promising, capturing coarse dynamics at low temporal/spatial resolution and refining details progressively (*e.g.*, Motion-I2V (Shi et al. 2024b)). For further

scalability, memory-efficient architectures (*e.g.*, sparse attention, recurrent latent modules) can reduce long-range modeling cost, and large-scale self-supervised pretraining on datasets such as WebVid-10M (Bain et al. 2021) and Panda-70M (Chen et al. 2024c) can improve generalization and reduce reliance on task-specific tuning.

Overall, advances in hybrid architectures and efficient long-range modeling are likely to be central to practical, high-quality, and controllable long-video generation for applications such as storytelling, simulation, and animation.

8 Conclusion

This paper presents a comprehensive survey of controllable video generation based on foundational generative models. First, we introduce the core theoretical foundations, including GANs, VAEs, denoising diffusion probabilistic models, flow-based models, and autoregressive architectures, along with representative video generative models. Then, we propose a structured taxonomy that categorizes existing controllable generation methods by their conditioning signals beyond text. Next, we review representative techniques for incorporating novel conditions into video generation pipelines, tracing their development across different theoretical paradigms and model designs. We further synthesize prior research by examining its progression from core principles to technical innovations and implementation strategies. Additionally, we highlight real-world applications where controllable video generation has demonstrated its practical impact, emphasizing its relevance and future promise within the broader AIGC ecosystem. Through this survey, we aim to present a holistic understanding of the field’s current landscape and provide insights that inform future directions in controllable video synthesis research.

Data Availability Statements. All the data used in this paper can be downloaded from the Internet, and the authorization has been obtained, including civitai ([civ 2023](#)).

References

- (2023) civitai. <https://civitai.com/models/443821/cyberrealistic-pony>

- (2025) Laion-coco. Accessed October 22, 2023 [Online] <https://laion.ai/blog/laion-coco/>
- (2025) Pexels. royalty-free stock footage website, 2014, URL <https://www.pexels.com/>
- Agarwal N, Ali A, Bala M, et al (2025) Cosmos world foundation model platform for physical ai. arXiv:250103575
- Akerman R, Feng H, Black MJ, et al (2025) Inter-dyn: Controllable interactive dynamics with video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 12467–12479
- Aldausari N, Sowmya A, Marcus N, et al (2022) Video generative adversarial networks: a review. ACM Computing Surveys
- Bahmani S, Skorokhodov I, Rong V, et al (2024) 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In: CVPR
- Bahmani S, Skorokhodov I, Qian G, et al (2025a) Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 22875–22889
- Bahmani S, Skorokhodov I, Siarohin A, et al (2025b) Vd3d: Taming large video diffusion transformers for 3d camera control. arXiv:240712781
- Bai J, Xia M, et al (2025) Recammaster: Camera-controlled generative rendering from a single video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 14834–14844
- Bain M, Nagrani A, Varol G, et al (2021) Frozen in time: A joint video and image encoder for end-to-end retrieval. In: ICCV
- Bar-Tal O, Chefer H, Tov O, et al (2024) Lumiere: A space-time diffusion model for video generation. In: SIGGRAPH Asia
- Bharadhwaj H, Dwibedi D, Gupta A, et al (2024) Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. In: arXiv.org, <https://doi.org/10.48550/arXiv.2409.16283>
- Bi X, Lu J, Liu B, et al (2025) Customttt: Motion and appearance customized video generation via test-time training. In: AAAI
- Bian W, Huang Z, Shi X, et al (2025a) Gsdit: Advancing video generation with dynamic 3d gaussian fields through efficient dense 3d point tracking. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 21717–21727, <https://doi.org/10.1109/CVPR52734.2025.02023>
- Bian Y, Zeng A, et al (2025b) Motioncraft: Crafting whole-body motion with plug-and-play multimodal controls. In: AAAI
- Biyyala V, Kathuria BC, Li J, et al (2025) Sstem: Advanced metrics for evaluating semantic spatial and temporal aspects in video editing. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV) Workshops, pp 259–268
- Blattmann A, Dockhorn T, Kulal S, et al (2023) Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv:231115127
- Caesar H, Bankiti V, Lang AH, et al (2019) nuscenes: A multimodal dataset for autonomous driving. In: CVPR, <https://doi.org/10.1109/cvpr42600.2020.01164>
- Cao C, Zhou J, et al (2025) Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. Association for Computing Machinery, New York, NY, USA, SA Conference Papers '25, <https://doi.org/10.1145/3757377.3763842>, URL <https://doi.org/10.1145/3757377.3763842>
- Cao Z, Hidalgo Martinez G, Simon T, et al (2019) Openpose: Realtime multi-person 2d pose estimation using part affinity fields. TPAMI

- Chang D, Shi Y, Gao Q, et al (2023) Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. arXiv:231112052
- Chang H, Zhang H, Jiang L, et al (2022) Maskgit: Masked generative image transformer. In: CVPR, <https://doi.org/10.1109/CVPR52688.2022.01103>
- Chefer H, Zada S, Paiss R, et al (2024) Still-moving: Customized video generation without customized video data. ACM TOG
- Chen B, Chen S, Yue Z, et al (2025a) G-ubs: Towards robust understanding of implicit feedback via group-aware user behavior simulation. arXiv preprint arXiv:250805709
- Chen B, Wang Z, Yue Z, et al (2025b) Videochatm1: Collaborative policy planning for video understanding via multi-agent reinforcement learning. arXiv preprint arXiv:251119524
- Chen D, Liao J, Yuan L, et al (2017) Coherent online video style transfer. In: ICCV
- Chen H, Xie W, Vedaldi A, et al (2020) Vggsound: A large-scale audio-visual dataset. In: ICASSP
- Chen H, Wang X, Zeng G, et al (2023a) Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. arXiv:231100990
- Chen H, Xia M, et al (2023b) Videocrafter1: Open diffusion models for high-quality video generation. arXiv:231019512
- Chen H, Wang X, Zhang Y, et al (2024a) Disenstudio: Customized multi-subject text-to-video generation with disentangled spatial control. In: Proceedings of the 32nd ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, MM '24, p 3637–3646, <https://doi.org/10.1145/3664647.3680637>, URL <https://doi.org/10.1145/3664647.3680637>
- Chen Q, Ma Y, Wang H, et al (2024b) Follow-your-canvas: Higher-resolution video outpainting with extensive content generation. arXiv:240901055
- Chen RT, Rubanova Y, Bettencourt J, et al (2018) Neural ordinary differential equations. NeurIPS
- Chen S, Chen B, Yu C, et al (2025c) When top-ranked recommendations fail: Modeling multi-granular negative feedback for explainable and robust video recommendation. arXiv preprint arXiv:251118700
- Chen TS, Lin CH, Tseng HY, et al (2023c) Motion-conditioned diffusion model for controllable video synthesis. arXiv:230414404
- Chen TS, Siarohin A, Menapace W, et al (2024c) Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In: CVPR
- Chen W, Ji Y, Wu J, et al (2023d) Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. arXiv:230513840
- Chen X, Chen Z, Song Y (2025d) Transanimate: Taming layer diffusion to generate rgba video. arXiv preprint arXiv:250317934
- Chen Y, Men Y, Yao Y, et al (2025e) Perception-as-control: Fine-grained controllable image animation with 3d-aware motion representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 14380–14389
- Chen Z, Cao J, Chen Z, et al (2025f) Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In: AAAI
- Chi C, Xu Z, Feng S, et al (2024a) Diffusion policy: Visuomotor policy learning via action diffusion. IJRR
- Chi C, Xu Z, Pan C, et al (2024b) Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In: Robotics: Science and Systems, <https://doi.org/10.48550/arXiv.2402.10329>
- Clark J, Mirchandani S, Sadigh D, et al (2025) Action-free reasoning for policy generalization. arXiv:250203729

- Deng H, Pan T, Diao H, et al (2024) Autoregressive video generation without vector quantization. arXiv:241214169
- Deng Y, Tang F, Dong W, et al (2021) Arbitrary video style transfer via multi-channel correlation. In: AAAI
- Ding M, Yang Z, Hong W, et al (2021) Cogview: Mastering text-to-image generation via transformers. In: NeurIPS
- Dinh L, Krueger D, Bengio Y (2014) Nice: Non-linear independent components estimation. arXiv:14108516
- Dinh L, Sohl-Dickstein J, Bengio S (2016) Density estimation using real nvp. arXiv:160508803
- Esser P, Chiu J, Atighchian P, et al (2023) Structure and content-guided video synthesis with diffusion models. In: ICCV
- Feng K, Ma Y, Wang B, et al (2025a) Dit4edit: Diffusion transformer for image editing. In: AAAI
- Feng K, Ma Y, Zhang X, et al (2025b) Follow-your-instruction: A comprehensive mllm agent for world data synthesis. arXiv preprint arXiv:250805580
- Feng W, Qi T, Liu J, et al (2025c) I2vcontrol: Disentangled and unified video motion synthesis control. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 14051–14060
- Fu X, Liu X, et al (2024) 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. arXiv:241207759
- Fu X, Wang X, Liu X, et al (2025) Learning video generation for robotic manipulation with collaborative trajectory control. arXiv:250601943
- Gafni O, Polyak A, Ashual O, et al (2022) Make-a-scene: Scene-based text-to-image generation with human priors. In: ECCV, <https://doi.org/10.48550/arXiv.2203.13131>
- Gao R, Chen K, Xie E, et al (2023) Magicdrive: Street view generation with diverse 3d geometry control. arXiv:231002601
- Gao R, Chen K, Xiao B, et al (2025) Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 28135–28144
- Gao S, Yang J, Chen L, et al (2024) Vista: A generalizable driving world model with high fidelity and versatile controllability. In: NeurIPS, <https://doi.org/10.48550/arXiv.2405.17398>
- Gao W, et al (2020) Fast video multi-style transfer. In: WACV
- Gemmeke J, et al (2017) Audio set: An ontology and human-labeled dataset for audio events. In: ICASSP, <https://doi.org/10.1109/ICASSP.2017.7952261>
- Geng D, et al (2025) Motion prompting: Controlling video generation with motion trajectories. In: CVPR
- Girdhar R, et al (2023) Emu video: Factorizing text-to-video generation by explicit image conditioning. arXiv:231110709
- Gong Y, Song Y, Li Y, et al (2025) Relation-adapter: Learning and transferring visual relation with diffusion transformers. arXiv preprint arXiv:250602528
- Gong Y, et al (2023) Talecrafter: Interactive story visualization with multiple characters. arXiv:230518247
- Goodfellow IJ, et al (2014) Generative adversarial nets. NeurIPS
- Gu Z, et al (2025) Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In: SIGGRAPH Conference Papers, <https://doi.org/10.1145/3721238.3730607>
- Güler RA, et al (2018) Densepose: Dense human pose estimation in the wild. In: CVPR
- Guo C, et al (2022) Generating diverse and natural 3d human motions from text. In: CVPR

- Guo H, Zeng B, Song Y, et al (2025) Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. arXiv preprint arXiv:250115891
- Guo J, et al (2024a) Liveportrait: Efficient portrait animation with stitching and retargeting control. arXiv:240703168
- Guo X, et al (2024b) I2v-adapter: A general image-to-video adapter for diffusion models. In: SIGGRAPH
- Guo Y, et al (2024c) Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: ICLR
- Guo Y, et al (2024d) Sparsectrl: Adding sparse controls to text-to-video diffusion models. In: ECCV
- Haji-Ali M, et al (2025) Av-link: Temporally-aligned diffusion features for cross-modal audio-video generation. In: ICCV
- He H, et al (2024a) Cameractrl: Enabling camera control for text-to-video generation. arXiv:240402101
- He H, et al (2025) Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. arXiv:250310592
- He X, et al (2024b) Id-animator: Zero-shot identity-preserving human video generation. arXiv:240415275
- He Y, et al (2022) Latent video diffusion models for high-fidelity long video generation. arXiv:221113221
- Ho J, et al (2020) Denoising diffusion probabilistic models. NeurIPS
- Ho J, et al (2022) Imagen video: High definition video generation with diffusion models. arXiv:221002303
- Hong FT, et al (2025) Audio-visual controlled video diffusion with masked selective state spaces modeling for natural talking head generation. In: ICCV
- Hu L (2024) Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In: CVPR
- Hu L, et al (2025a) Animate anyone 2: High-fidelity character image animation with environment affordance. arXiv:250206145
- Hu T, et al (2025b) Harmony: Harmonizing audio and video generation through cross-task synergy. arXiv preprint arXiv:251121579
- Huang CP, et al (2025a) Videomage: Multi-subject and motion customization of text-to-video diffusion models. In: CVPR
- Huang H, et al (2017) Real-time neural style transfer for videos. In: CVPR
- Huang T, Zheng W, et al (2025b) Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. ACM TOG <https://doi.org/10.1145/3763330>
- Huang Y, et al (2025c) Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. arXiv:250104698
- Huang Z, et al (2024) Lvcd: reference-based lineart video colorization with diffusion models. ACM TOG
- Jain Y, et al (2024) Peekaboo: Interactive video generation via masked-diffusion. In: CVPR
- Jeong H, et al (2024) Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In: CVPR
- Jeong H, et al (2025a) Dreammotion: Space-time self-similar score distillation for zero-shot video editing. In: ECCV
- Jeong H, et al (2025b) Track4gen: Teaching video diffusion models to track points improves video generation. In: CVPR
- Jiang J, et al (2024a) Dive: Dit-based video generation with enhanced control. arXiv:240901595
- Jiang Y, Gu Y, Song Y, et al (2025a) Personalized vision via visual in-context learning. arXiv

- preprint arXiv:250925172
- Jiang Y, et al (2024b) Videobooth: Diffusion-based video generation with image prompts. In: CVPR
- Jiang Z, et al (2025b) Vace: All-in-one video creation and editing. In: ICCV
- Jin X, Xu Z, Ou M, et al (2024) Alignment is all you need: A training-free augmentation strategy for pose-guided video generation. arXiv preprint arXiv:240816506
- Ju X, Gao Y, Zhang Z, et al (2024) Miradata: A large-scale video dataset with long durations and structured captions. In: NeurIPS, <https://doi.org/10.48550/arXiv.2407.06358>
- Ju X, Ye W, Liu Q, et al (2025) Fulldit: Multi-task video generative foundation model with full attention. arXiv:250319907
- Karras J, Holynski A, Wang TC, et al (2023) Dreampose: Fashion image-to-video synthesis via stable diffusion. In: ICCV
- Ki T, Min D, Chae G (2024) Float: Generative motion latent flow matching for audio-driven talking portrait. arXiv:241201064
- Kim SW, Philion J, Torralba A, et al (2021) Drivagan: Towards a controllable high-quality neural simulation. In: CVPR
- Kingma DP, Dhariwal P (2018) Glow: Generative flow with invertible 1x1 convolutions. NeurIPS
- Kingma DP, Welling M, et al (2013) Auto-encoding variational bayes
- Ko PC, Mao J, Du Y, et al (2023) Learning to act from actionless videos through dense correspondences. In: ICCV, <https://doi.org/10.48550/arXiv.2310.08576>
- Kong W, Tian Q, Zhang Z, et al (2024) Hunyuan-video: A systematic framework for large video generative models. arXiv:241203603
- Koroglu M, Caselles-Dupré H, Jeanneret G, et al (2025) Onlyflow: Optical flow based motion conditioning for video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp 6280–6290
- Kuang Z, Cai S, et al (2024) Collaborative video diffusion: Consistent multi-video generation with camera control. In: Globerson A, Mackey L, Belgrave D, et al (eds) Advances in Neural Information Processing Systems, vol 37. Curran Associates, Inc., pp 16240–16271, <https://doi.org/10.5220/079017-0518>, URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1d49235669869ab737c1da9d64b7c769-Paper-Conference.pdf
- Lee SH, Oh G, Byeon W, et al (2022) Sound-guided semantic video generation. In: ECCV, <https://doi.org/10.48550/arXiv.2204.09273>
- Li H, Li Y, Yang Y, et al (2024a) Dispose: Disentangling pose guidance for controllable human image animation. arXiv:241209349
- Li H, Qiu H, Zhang S, et al (2025a) Personavideo: High id-fidelity video customization without dynamic and semantic degradation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 19406–19416
- Li Q, Xing Z, Wang R, et al (2025b) Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 12112–12123
- Li R, Zhao J, Zhang Y, et al (2023) Finedance: A fine-grained choreography dataset for 3d full body dance generation. In: ICCV
- Li S, Gao Y, Sadigh D, et al (2025c) Unified video action model. arXiv:250300200
- Li T, Zheng G, Jiang R, et al (2025d) Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 28785–28796

- Li X, Jia X, Wang Q, et al (2024b) Motrans: Customized motion transfer with text-driven video diffusion models. Association for Computing Machinery, New York, NY, USA, MM '24, p 3421–3430, <https://doi.org/10.1145/3664647.3680718>, URL <https://doi.org/10.1145/3664647.3680718>
- Li X, Wang J, Cheng Y, et al (2025e) Towards high-fidelity 3d talking avatar with personalized dynamic texture. In: CVPR
- Li X, Wu C, Yang Z, et al (2025f) Driverse: Navigation world model for driving simulation via multimodal trajectory prompting and motion alignment
- Li X, Xue H, Ren P, et al (2025g) Difueraser: A diffusion model for video inpainting. arXiv:250110018
- Lian L, Shi B, Yala A, et al (2023) Llm-grounded video diffusion models. arXiv:230917444
- Liang F, Ma H, He Z, et al (2025) Movie weaver: Tuning-free multi-concept video personalization with anchored prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 13146–13156
- Liang J, Fan Y, Zhang K, et al (2024) Movideo: Motion-aware video generation with diffusion model. In: ECCV
- Liao X, Zeng X, Wang L, et al (2025) Motionagent: Fine-grained controllable video generation via motion field agent. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 11305–11316
- Lin B, Ge Y, Cheng X, et al (2024a) Open-sora plan: Open-source large video generation model. arXiv:241200131
- Lin B, Yu Y, Ye J, et al (2024b) Takin-ada: Emotion controllable audio-driven animation with canonical and landmark loss optimization. arXiv:241014283
- Lin G, Jiang J, Liang C, et al (2025a) Cyber-host: A one-stage diffusion framework for audio-driven talking body generation. In: ICLR
- Lin G, Jiang J, Yang J, et al (2025b) Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. arXiv:250201061
- Lin H, Cho J, Zala A, et al (2024c) Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. arXiv:240409967
- Lin Z, Liu W, Chen C, et al (2025c) Stiv: Scalable text and image conditioned video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 16249–16259
- Ling L, Sheng Y, Tu Z, et al (2023) Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: CVPR, <https://doi.org/10.1109/CVPR52733.2024.02092>
- Ling P, Bu J, Zhang P, et al (2025) Motionclone: Training-free motion cloning for controllable video generation. In: The Thirteenth International Conference on Learning Representations
- Lipman Y, Chen RT, Ben-Hamu H, et al (2022) Flow matching for generative modeling. arXiv:221002747
- Liu B, Zhu Y, Gao C, et al (2023a) Libero: Benchmarking knowledge transfer for lifelong robot learning. In: NeurIPS, <https://doi.org/10.48550/arXiv.2306.03310>
- Liu FL, Fu H, et al (2025a) Sketchvideo: Sketch-based video generation and editing. In: CVPR
- Liu G, Xia M, Zhang Y, et al (2024a) Stylecrafter: Taming artistic video diffusion with reference-augmented adapter learning. ACM Trans Graph 43(6). <https://doi.org/10.1145/3687975>, URL <https://doi.org/10.1145/3687975>
- Liu H, Zhu Z, Becherini G, et al (2023b) Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In: CVPR, <https://doi.org/10.1109/CVPR52733.2024.02092>

- Liu J, Han J, Yan B, et al (2025b) Infinitystar: Unified spacetime autoregressive modeling for visual generation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems, URL <https://openreview.net/forum?id=JcEqP4aPmb>
- Liu K, Shao L, Lu S (2024b) Novel view extrapolation with video diffusion priors. arXiv preprint arXiv:241114208
- Liu K, Li W, Chen L, et al (2025c) Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. arXiv preprint arXiv:250323377
- Liu L, Liu Q, Qian S, et al (2024c) Text-animator: Controllable visual text video generation. arXiv:240617777
- Liu L, Ma T, Li B, et al (2025d) Phantom: Subject-consistent video generation via cross-modal alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 14951–14961
- Liu S, Ren Z, Gupta S, et al (2024d) Physgen: Rigid-body physics-grounded image-to-video generation. In: ECCV
- Loftsdottir D, Guzdial M (2022) Sketchbetween: Video-to-video synthesis for sprite animation via sketches. In: FDG
- Long Z, Zheng M, Feng K, et al (2025) Follow-your-shape: Shape-aware image editing via trajectory-guided region control. arXiv preprint arXiv:250808134
- Low C, Wang W, Katyal C (2025) Ovi: Twin backbone cross-modal fusion for audio-video generation. arXiv preprint arXiv:251001284
- Lu H, Wu X, Wang S, et al (2024) Seeing beyond views: Multi-view driving scene video generation with holistic attention. arXiv:241203520
- Luo GY, Luo Z, Gosselin A, et al (2025a) Ctrl-v: Higher fidelity autonomous vehicle video generation with bounding-box controlled object motion. TMLR
- Luo Y, Rong Z, Wang L, et al (2025b) Dreamactor-m1: Holistic, expressive and robust human image animation with hybrid guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 11036–11046
- Ma B, Gao H, Deng H, et al (2025a) You see it, you got it: Learning 3d creation on pose-free videos at scale. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 2016–2029
- Ma E, Zhou L, Tang T, et al (2024a) Unleashing generalization of end-to-end autonomous driving with controllable long video generation. arXiv:240601349
- Ma G, Huang H, Yan K, et al (2025b) Step-video-t2v technical report: The practice, challenges, and future of video foundation model. arXiv:250210248
- Ma WDK, Lewis JP, Kleijn WB (2024b) Trailblazer: Trajectory control for diffusion-based video generation. In: SIGGRAPH Asia
- Ma Y, He Y, Cun X, et al (2024c) Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: AAAI
- Ma Y, Liu H, Wang H, et al (2024d) Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In: SIGGRAPH Asia
- Ma Y, Cun X, Liang S, et al (2025c) Magicstick: Controllable video editing via control handle transformations. In: WACV
- Ma Y, Feng K, Zhang X, et al (2025d) Follow-your-creation: Empowering 4d creation through video inpainting. arXiv:250604590
- Ma Y, He Y, Wang H, et al (2025e) Follow-your-click: Open-domain regional image animation via motion prompts. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 6018–6026
- Ma Z, Zhang Y, Jia G, et al (2025f) Efficient Diffusion Models: A Comprehensive Survey From Principles to Practices .

- IEEE Transactions on Pattern Analysis & Machine Intelligence 47(09):7506–7525. <https://doi.org/10.1109/TPAMI.2025.3569700>, URL <https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3569700>
- Ma Z, Zhou D, Wang XS, et al (2025g) Magic-me: Identity-specific video customized diffusion. In: Del Bue A, Canton C, Pont-Tuset J, et al (eds) Computer Vision – ECCV 2024 Workshops. Springer Nature Switzerland, Cham, pp 19–37
- Mei J, Hu T, Yang X, et al (2024) Dreamforge: Motion-aware autoregressive video generation for multi-view driving scenes. arXiv:240904003
- Melnik A, Ljubljancic M, Lu C, et al (2024) Video diffusion models: A survey. arXiv:240503150
- Meng R, Zhang X, Li Y, et al (2025a) Echomimicv2: Towards striking, simplified, and semi-body human animation. In: CVPR
- Meng Y, Ouyang H, Wang H, et al (2025b) Anidoc: Animation creation made easier. In: CVPR
- Moher D, Liberati A, Tetzlaff J, et al (2009) Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Bmj* 339
- Ni H, Egger B, et al. (2024) Ti2v-zero: Zero-shot image conditioning for text-to-video diffusion models. In: CVPR
- Niu M, Cun X, Wang X, et al (2024) Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In: ECCV
- Van den Oord A, et al (2016) Conditional image generation with pixelcnn decoders
- Pang Y, Zhu B, Lin B, et al (2024) Dreamdance: Animating human images by enriching 3d geometry cues from 2d poses. arXiv:241200397
- Park GY, et al (2025) Spectral motion alignment for video motion transfer using diffusion models. In: AAAI
- Peebles W, Xie S (2023) Scalable diffusion models with transformers. In: ICCV
- Peng B, et al (2024) Controlnext: Powerful and efficient control for image and video generation. arXiv:240806070
- Pont-Tuset J, et al (2017) The 2017 davis challenge on video object segmentation. arXiv:170400675
- list not fully provided A (2025) Motionflow: Learning implicit motion flow for complex camera trajectory control in video generation. OpenReview Published on 16 Mar 2025
- PYang S, et al (2022) Vtoonify: Controllable high-resolution portrait video style transfer. ACM TOG
- Radford A, et al (2021) Learning transferable visual models from natural language supervision. In: ICML
- Ramesh A, et al (2021) Zero-shot text-to-image generation. In: PMLR
- Rao A, et al (2020) A unified framework for shot type classification based on subject centric lens. In: ECCV
- Ren W, et al (2024a) Consisti2v: Enhancing visual consistency for image-to-video generation. arXiv:240204324
- Ren X, Shen T, et al (2025) Gen3c: 3d-informed world-consistent video generation with precise camera control. In: CVPR
- Ren Y, et al (2024b) Customize-a-video: One-shot motion customization of text-to-video diffusion models. In: ECCV
- Ruan L, et al (2023) Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In: CVPR
- Russell L, et al (2025) Gaia-2: A controllable multi-view generative world model for autonomous driving. arXiv:250320523

- Schneider S, Baevski A, et al (2019) wav2vec: Unsupervised pre-training for speech recognition. arXiv:190405862
- Schuhmann C, et al (2021) Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv:211102114
- Shen C, Gan Y, Chen C, et al (2024) Decouple content and motion for conditional image-to-video generation. In: AAAI
- Shen F, et al (2025) Long-term talking face generation via motion-prior conditional diffusion model. arXiv:250209533
- Shi S, et al (2025a) Motionstone: Decoupled motion intensity modulation with diffusion transformer for image-to-video generation. In: CVPR
- Shi W, Song Y, Zhang D, et al (2024a) Fonts: Text rendering with typography and style controls. arXiv preprint arXiv:241200136
- Shi W, Song Y, Rao Z, et al (2025b) Wordcon: Word-level typography control in scene text rendering. arXiv preprint arXiv:250621276
- Shi X, et al (2024b) Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In: SIGGRAPH
- Siarohin A, Woodford OJ, Ren J, et al (2021) Motion representations for articulated animation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13653–13662
- Song C, Yang Y, Zhao T, et al (2025a) Worldforge: Unlocking emergent 3d/4d generation in video diffusion model via training-free guidance. arXiv preprint arXiv:250915130
- Song K, et al (2025b) Llama learns to direct: Directorllm for human-centric video generation. In: BMVC
- Song Q, et al (2024a) Univst: A unified framework for training-free localized video style transfer. arXiv:241020084
- Song Y (2022) Cliptexture: Text-driven texture synthesis. In: Proceedings of the 30th ACM International Conference on Multimedia, pp 5468–5476
- Song Y, Zhang Y (2022) Clipfont: Text guided vector wordart generation. In: BMVC, p 543
- Song Y, Shao X, Chen K, et al (2023) Clipvg: Text-guided image manipulation using differentiable vector graphics. In: Proceedings of the AAAI conference on artificial intelligence, pp 2312–2320
- Song Y, Huang S, Yao C, et al (2024b) Process-painter: Learn painting process from sequence data. arXiv preprint arXiv:240606062
- Song Y, Liu X, Shou MZ (2024c) Diffsim: Taming diffusion models for evaluating visual similarity. arXiv preprint arXiv:241214580
- Song Y, Chen D, Shou MZ (2025c) Layer-tracer: Cognitive-aligned layered svg synthesis via diffusion transformer. arXiv preprint arXiv:250201105
- Song Y, Liu C, Shou MZ (2025d) Makeanything: Harnessing diffusion transformers for multi-domain procedural sequence generation. arXiv preprint arXiv:250201572
- Song Y, Liu C, Shou MZ (2025e) Omniconsistency: Learning style-agnostic consistency from paired stylization data. arXiv preprint arXiv:250518445
- Sun W, Tu RC, Liao J, et al (2024) Diffusion model-based video editing: A survey. arXiv:240707111
- Sun X, Zhang L, Zhu H, et al (2025) Vividtalk: One-shot audio-driven talking head generation based on 3d hybrid prior. In: 3DV
- Tan S, Ji B, Bi M, et al (2024a) Edtalk: Efficient disentanglement for emotional talking head synthesis. In: ECCV
- Tan W, Lin C, et al (2024b) Svp: Style-enhanced vivid portrait talking head diffusion model. arXiv:240903270

- Tan Z, et al (2025) Ominicontrol: Minimal and universal control for diffusion transformer. In: ICCV
- Teng H, et al (2025) Magi-1: Autoregressive video generation at scale. arXiv:250513211
- Tian J, et al (2025) Extrapolating and decoupling image-to-video generation models: Motion modeling is easier than you think. In: CVPR
- Tian K, et al (2024a) Visual autoregressive modeling: Scalable image generation via next-scale prediction. In: NeurIPS
- Tian L, Wang Q, Zhang B, et al (2024b) Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In: ECCV
- Tu Y, et al (2025) Videoanydoor: High-fidelity video object insertion with precise motion control. In: SIGGRAPH
- Wan T, et al (2025) Wan: Open and advanced large-scale video generative models. arXiv:250320314
- Wan Z, et al (2024) Dragentity: Trajectory guided video generation using entity and positional relationships. In: ACM MM
- Wang B, et al (2025a) This&that: Language-gesture controlled video generation for robot planning
- Wang C, et al (2024a) V-express: Conditional dropout for progressive training of portrait video generation. arXiv:240602511
- Wang C, et al (2025b) Dreamvideo: High-fidelity image-to-video generation with image retention and text guidance. In: ICASSP
- Wang D, et al (2025c) Universe-1: Unified audio-video generation via stitching of experts. arXiv:250906155
- Wang H, et al (2025d) Levitor: 3d trajectory oriented image-to-video synthesis. In: CVPR
- Wang J, et al (2024b) Boximator: Generating rich and controllable motions for video synthesis. arXiv:240201566
- Wang J, et al (2024c) Cove: Unleashing the diffusion feature correspondence for consistent video editing. arXiv:240608850
- Wang J, et al (2024d) Omnitokenizer: A joint image-video tokenizer for visual generation. NeurIPS
- Wang Q, Shi X, Li B, et al (2025e) Multishotmaster: A controllable multi-shot video generation framework. arXiv preprint arXiv:251203041
- Wang Q, et al (2025f) Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation
- Wang T, et al (2024e) Disco: Disentangled control for realistic human dance generation. In: CVPR
- Wang TH, Cheng YC, Lin CH, et al (2019) Point-to-point video generation. In: ICCV
- Wang W, et al (2024f) Mvoc: a training-free multiple video object composition method with diffusion models. arXiv:240615829
- Wang W, et al (2025g) Dualreal: Adaptive joint training for lossless identity-motion fusion in video customization. In: ICCV
- Wang X, et al (2023a) Videocomposer: Compositional video synthesis with motion controllability. NeurIPS
- Wang X, et al (2023b) Videolcm: Video latent consistency model. arXiv:231209109
- Wang X, et al (2024g) A recipe for scaling up text-to-video generation with text-free videos. In: CVPR
- Wang Y, et al (2020) Imaginator: Conditional spatio-temporal gan for video generation. In: WACV
- Wang Y, et al (2025h) Survey of video diffusion models: Foundations, implementations, and applications. arXiv:250416081
- Wang YQ, et al (2023c) Driving into the future: Multiview visual forecasting and planning with

- world model for autonomous driving. In: CVPR
- Wang Z, Yuan Z, Wang X, et al (2024h) Motionctrl: A unified and flexible motion controller for video generation. arXiv:231203641
- Wang Z, Zhao H, Zhou Q, et al (2025i) Diffdecompose: Layer-wise decomposition of alpha-composited images via diffusion transformers. arXiv preprint arXiv:250521541
- Wang Z, et al (2024i) Customvideo: Customizing text-to-video generation with multiple subjects. arXiv:240109962
- Wei H, et al (2024a) Aniprtrait: Audio-driven synthesis of photorealistic portrait animation. arXiv:240317694
- Wei Y, et al (2024b) Dreamvideo: Composing your dream videos with customized subject and motion. In: CVPR
- Wei Y, et al (2025) Dreamrelation: Relation-centric video customization. In: ICCV
- Wen Y, et al (2024) Panacea: Panoramic and controllable video generation for autonomous driving. In: CVPR
- Wu C, et al (2021) NÜwa: Visual synthesis pre-training for neural visual world creation. In: ECCV
- Wu J, et al (2024a) Motionbooth: Motion-aware customized text-to-video generation. In: NeurIPS
- Wu JZ, et al (2023) Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: ICCV
- Wu S, et al (2025a) Any2caption: Interpreting any condition to caption for controllable video generation. arXiv:250324379
- Wu T, et al (2025b) Customcrafter: Customized video generation with preserving motion and concept composition abilities. In: AAAI
- Wu T, et al (2025c) Video world models with long-term spatial memory. arXiv:250605284
- Wu W, et al (2024b) Draganything: Motion control for anything using entity representation. In: ECCV
- Wu YS, et al (2025d) Motionmatcher: Cinematic motion customization of text-to-video diffusion models via motion feature matching. In: ICCV
- Xiao S, et al (2025a) Omnigen: Unified image generation. In: CVPR
- Xiao Z, et al (2025b) Trajectory attention for fine-grained video motion control. In: ICLR
- Xie K, et al (2025) Videopanda: Video panoramic diffusion with multi-view attention. arXiv:250411389
- Xie L, et al (2022) Vfhq: A high-quality dataset and benchmark for video face super-resolution. In: CVPRW
- Xie Y, et al (2024) X-portrait: Expressive portrait animation with hierarchical motion attention. In: SIGGRAPH'24
- Xing J, et al (2024a) Dynamiccrafter: Animating open-domain images with video diffusion priors. In: ECCV
- Xing J, et al (2024b) Make-your-video: Customized video generation using textual and structural guidance. TVCG
- Xing J, et al (2024c) Tooncrafter: Generative cartoon interpolation. ACM TOG
- Xing J, et al (2025a) Motioncanvas: Cinematic shot design with controllable image-to-video generation. In: SIGGRAPH'25
- Xing Z, et al (2024d) A survey on video diffusion models. ACM Computing Surveys
- Xing Z, et al (2025b) Aid: Adapting image2video diffusion models for instruction-guided video prediction. In: ICCV
- Xu D, et al (2024a) Camco: Camera-controllable 3d-consistent image-to-video generation. arXiv:240602011

- Xu J, et al (2024b) Easyanimate: A high-performance long video generation method based on transformer architecture. arXiv:240518991
- Xu M, et al (2024c) Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. arXiv:240608801
- Xu S, Chen G, et al (2024d) Vasa-1: Lifelike audio-driven talking faces generated in real time. NeurIPS
- Xu Z, et al (2024e) Magicanimate: Temporally consistent human image animation using diffusion model. In: CVPR
- Xu Z, et al (2025) Hunyuanportrait: Implicit condition control for enhanced portrait animation. In: CVPR
- Xue H, et al (2021) Advancing high-resolution video-language representation with large-scale video transcriptions. In: CVPR
- Xue J, Wang H, Tian Q, et al (2024) Follow-your-pose v2: Multiple-condition guided character image animation for stable pose control. arXiv e-prints pp arXiv-2406
- Yang C, Wang Z, Zhu X, et al (2018) Pose guided human video generation. In: ECCV
- Yang S, Zhou Y, Liu Z, et al (2024a) Fresco: Spatial-temporal correspondence for zero-shot video translation. In: CVPR
- Yang S, et al (2023a) Rerender a video: Zero-shot text-guided video-to-video translation. In: SIGGRAPH Asia
- Yang S, et al (2023b) Styleganex: Stylegan-based manipulation beyond cropped aligned faces. In: ICCV
- Yang S, et al (2024b) Direct-a-video: Customized video generation with user-directed camera movement and object motion. SIGGRAPH
- Yang S, et al (2025a) Megactor-sigma: Unlocking flexible mixed-modal control in portrait animation with diffusion transformer. In: AAAI
- Yang X, Zhu L, Fan H, et al (2025b) Videograin: Modulating space-time attention for multi-grained video editing. In: ICLR
- Yang Y, Fan L, Lin Z, et al (2025c) Layeranimate: Layer-level control for animation. In: ICCV
- Yang Z, et al (2025d) Cogvideox: Text-to-video diffusion models with an expert transformer. ICLR
- Yariv G, et al (2024) Diverse and aligned audio-to-video generation via text-to-video model adaptation. In: AAAI
- Yariv G, et al (2025) Through-the-mask: Mask-based motion trajectories for image-to-video generation. In: CVPR
- Yazdani S, et al (2024) A comprehensive survey of image and video generative ai: Recent advances, variants, and applications
- Ye Z, et al (2025) Stylemaster: Stylize your video with artistic generation and translation. In: CVPR
- Yin T, et al (2025) From slow bidirectional to fast autoregressive video diffusion models. In: CVPR
- You M, Zhu Z, Liu H, et al (2024) Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. arXiv preprint arXiv:240515364
- Yu E, Blackburn-Matzen K, et al (2023) Videodoodles: Hand-drawn animations on videos with scene-aware canvases. ACM TOG
- Yu H, et al (2024) 4real: Towards photorealistic 4d scene generation via video diffusion models. In: NeurIPS
- Yu J, et al (2022) Scaling autoregressive models for content-rich text-to-image generation. In: TMLR
- YU M, Hu W, Xing J, et al (2025) Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. arXiv preprint arXiv:250305638

- Yu W, Xing J, et al (2025a) ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis . TPAMI
- Yu W, et al (2025b) Egosim: Egocentric exploration in virtual worlds with multi-modal conditioning. ICLR
- Yuan S, et al (2025) Identity-preserving text-to-video generation by frequency decomposition. In: CVPR
- Zeng Y, et al (2024) Make pixels dance: High-dynamic video generation. In: CVPR
- Zhang DJ, et al (2024a) Moonshot: Towards controllable video generation and editing with multimodal conditions. arXiv:240101827
- Zhang H, Chen K, Zhang Z, et al (2025a) Dualcamctrl: Dual-branch diffusion model for geometry-aware camera-controlled video generation. arXiv preprint arXiv:251123127
- Zhang J, Cai K, Fan Y, et al (2025b) MAT-agent: Adaptive multi-agent training optimization. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems, URL <https://openreview.net/forum?id=YDWRTYgR79>
- Zhang J, Cai K, Fan Y, et al (2025c) Cf-vlm:counterfactual vision-language fine-tuning. URL <https://arxiv.org/abs/2506.17267>, arXiv:2506.17267
- Zhang J, Cai K, Guo X, et al (2025d) Mm-cot:a benchmark for probing visual chain-of-thought reasoning in multimodal models. URL <https://arxiv.org/abs/2512.08228>, arXiv:2512.08228
- Zhang J, Cai K, Yang J, et al (2025e) Learning dynamics of vlm finetuning. URL <https://arxiv.org/abs/2510.11978>, arXiv:2510.11978
- Zhang J, Fan Y, Cai K, et al (2025f) Drdiff: Dynamic routing diffusion with hierarchical attention for breaking the efficiency-quality trade-off. URL <https://arxiv.org/abs/2509.02785>, arXiv:2509.02785
- Zhang J, Fan Y, Lin W, et al (2025g) GAM-agent: Game-theoretic and uncertainty-aware collaboration for complex visual reasoning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems, URL <https://openreview.net/forum?id=EKJhU5ioSo>
- Zhang J, Fan Y, Wen Z, et al (2025h) Tri-MARF: A tri-modal multi-agent responsive framework for comprehensive 3d object annotation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems, URL <https://openreview.net/forum?id=YGlbwfNWot>
- Zhang J, Guo X, Cai K, et al (2025i) Hybridtoken-vlm: Hybrid token compression for vision-language models. URL <https://arxiv.org/abs/2512.08240>, arXiv:2512.08240
- Zhang J, Huang Z, Fan Y, et al (2025j) KABB: Knowledge-aware bayesian bandits for dynamic expert coordination in multi-agent systems. In: Forty-second International Conference on Machine Learning, URL <https://openreview.net/forum?id=AKvy9a4jho>
- Zhang L, Rao A, Agrawala M (2023a) Adding conditional control to text-to-image diffusion models. In: ICCV
- Zhang L, Mo S, Zhang Y, et al (2024b) Audio-synchronized visual animation. In: ECCV
- Zhang M, Wu T, et al (2025k) Gendop: Auto-regressive camera trajectory generation as a director of photography. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 18229–18239
- Zhang R, Zhang M, Zhou J, et al (2025l) Mind-v: Hierarchical video generation for long-horizon robotic manipulation with rl-based physical alignment. arXiv preprint arXiv:251206628
- Zhang Y, Wei Y, Jiang D, et al (2023b) Controlvideo: Training-free controllable text-to-video generation. arXiv:230513077
- Zhang Y, Gu J, Wang LW, et al (2024c) Mimic-motion: High-quality human motion video generation with confidence-aware pose guidance. arXiv:240619680

- Zhang Y, Song Y, Liu J, et al (2024d) Ssr-encoder: Encoding selective subject representation for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8069–8078
- Zhang Y, Wei L, Zhang Q, et al (2024e) Stable-makeup: When real-world makeup transfer meets diffusion model. arXiv preprint arXiv:240307764
- Zhang Y, Yuan Y, Song Y, et al (2025m) Easycontrol: Adding efficient and flexible control for diffusion transformer. arXiv preprint arXiv:250307027
- Zhang Y, Zhang Q, Song Y, et al (2025n) Stable-hair: Real-world hair transfer via diffusion model. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 10348–10356
- Zhang Z, Li L, Ding Y, et al (2021) Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: CVPR
- Zhang Z, Long F, et al. (2024f) Trip: Temporal residual learning with image noise prior for image-to-video diffusion models. In: CVPR
- Zhang Z, Wu B, Wang X, et al (2024g) Avid: Any-length video inpainting with diffusion model. In: CVPR
- Zhang Z, Liao J, Li M, et al (2025o) Tora: Trajectory-oriented diffusion transformer for video generation. In: CVPR
- Zhang Z, Long F, Qiu Z, et al (2025p) Motionpro: A precise motion controller for image-to-video generation. In: CVPR
- Zhao G, Wang X, Zhu Z, et al (2025a) Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In: AAAI
- Zhao H, Liu X, Xu M, et al (2025b) Taste-rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation. In: CVPR
- Zhao J, Zhang H (2022) Thin-plate spline motion model for image animation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3657–3666
- Zhao M, Wang W, Chen T, et al (2024a) Ta2v: Text-audio guided video generation. TMM
- Zhao M, Zhu H, Xiang C, et al (2024b) Identifying and solving conditional image leakage in image-to-video diffusion model. In: Globerson A, Mackey L, Belgrave D, et al (eds) Advances in Neural Information Processing Systems, vol 37. Curran Associates, Inc., pp 30300–30326, <https://doi.org/10.52202/079017-0954>
- Zhao R, Gu Y, Wu JZ, et al (2025c) Motiondirector: Motion customization of text-to-video diffusion models. In: Leonardis A, Ricci E, Roth S, et al (eds) Computer Vision – ECCV 2024. Springer Nature Switzerland, Cham, pp 273–290
- Zheng G, Li T, et al (2024) Cami2v: Camera-controlled image-to-video diffusion model. arXiv:241015957
- Zhou H, Yu W, et al (2025) Holotime: Tampering video diffusion models for panoramic 4d scene generation. In: Proceedings of the 33rd ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, MM '25, p 9763–9772, <https://doi.org/10.1145/3746027.3755128>, URL <https://doi.org/10.1145/3746027.3755128>
- Zhou P, Wang L, Liu Z, et al (2024a) A survey on generative ai and llm for video generation, understanding, and streaming. arXiv:240416038
- Zhou T, Tucker R, Flynn J, et al (2018) Stereo magnification: Learning view synthesis using multiplane images. arXiv:180509817
- Zhou Y, Zhang R, Gu J, et al (2024b) Sugar: Subject-driven video customization in a zero-shot manner. arXiv:241210533
- Zhu C, Li K, Ma Y, et al (2024a) Instantswap: Fast customized concept swapping across sharp shape differences. arXiv preprint arXiv:241201197

Zhu H, Wu W, Zhu W, et al (2022) Celebvhq: A large-scale video facial attributes dataset. In: ECCV, <https://doi.org/10.48550/arXiv.2207.12393>

Zhu H, Wang Y, Zhou J, et al (2025a) Aether: Geometric-aware unified world modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 8535–8546

Zhu S, Chen JL, Dai Z, et al (2024b) Champ: Controllable and consistent human image animation with 3d parametric guidance. arXiv:240314781

Zhu T, Ren D, Wang Q, et al (2025b) Generative inbetweening through frame-wise conditions-driven video generation. In: CVPR

Zhuang S, Li K, Chen X, et al (2024) Vlogger: Make your dream a vlog. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 8806–8817

Zhuo J, Zhao X, et al (2023) Synthesizing videos from images for image-to-video adaptation. In: ACM MM