A Survey of Emerging Approaches and Advances in Video Generation

Elnaz Soleimani, Ghazaleh Khodabandelou

A Survey of Emerging Approaches and Advances in Video Generation

Elnaz Soleimani*, Ghazaleh Khodabandelou†

***Abstract*—The field of AI-driven video generation is evolving rapidly, with remarkable advancements achieved over the past two years. These developments have markedly enhanced the ability to transform human imagination into realistic visual content across various domains. This survey provides a comprehensive review of contemporary research in video generation, delving into foundational principles—including diverse generation strategies and key generative frameworks—as well as state-of-the-art models encompassing video synthesis, editing, and enhancement techniques. We present a comparative analysis of critical components such as core technologies, video quality attributes, hardware prerequisites, openness, and essential evaluation metrics and datasets. The survey concludes by discussing open challenges and emerging directions in the field, such as the role of LLMs (Large Language Model) and VLMs (Vision-Language Model) in advancing the sophistication of video generation frameworks. We intend for this survey to serve as a useful resource for researchers and practitioners, offering a structured overview of recent advancements and a clear depiction of the current landscape in video generation research.**

***Index Terms*—Generative AI, Video Generation, Diffusion Models, Flow Matching**

I. Introduction

THE task of Video Generation has recently attracted significant attention from both academic and industrial researchers. This heightened interest can be attributed to projections that the global media market will reach approximately $64.5 billion by 2031, with a compound annual growth rate of 14.8% from 2023 to 2031. This growth is propelled by the rising demand for high-quality video content, the expansion of video streaming services, and an increasing need for automated content generation [1]. Video is expected to become one of the dominant formats of communication across various domains -such as entertainment, marketing, and education- due to its engaging nature [2].

In recent years, video generation has made remarkable strides, accelerating the path toward achieving Artificial General Intelligence (AGI) and large-scale automatic generation of high-quality visual content. The intersection of breakthroughs in visual content generation and large models—such as Large Language Models (LLMs), Vision-Language Models (VLMs), and Large-Action Models (LAMs)— has accelerated video generation to new levels of creativity and efficiency. While LAMs have been briefly discussed in association with LLMs and VLMs in the context of video generation, they likely hold greater applicability in task-specific or action-driven video generation [3]–[5]. Specifically, LAMs are suited to usecases that necessitate the generation of coherent, contextually accurate sequences of actions. The integration of LAMs into video generation frameworks offers the potential to generate action-specific video content, enabling dynamic control over video outputs to adhere to predetermined action sequences. This capability is particularly pertinent in multimodal generation tasks, where the inclusion of structured action elements can significantly enhance realism and alignment with contextual requirements across domains such as simulation, robotics, education, and entertainment. A more comprehensive exploration of LAMs underscores their potential to advance sophisticated control mechanisms within video generation models, thereby establishing them as essential components in multimodal and action-centric video synthesis.

With the advent of diffusion models [6] surpassing Generative Adversarial Networks (GANs), coupled with major advancements in Natural Language Processing (NLP), a new era of video content creation has ushered. Video generation has been explored through various schemes including Text-to-Video(T2V), Image-to-Video(I2V), Video-to-Video(V2V), and Multi-modal video generation (X2V). Among these, text and image are considered the most prevalent modalities used for video generation. To leverage the advances in Text-to-Image (T2I) generation, T2V video generation is sometimes factorized into a two-stage process: First creating images through T2I generation, followed by an I2V generation process, using training free or few-shot [7], [8] learning methods. The growing interest in versatile and highly controllable video content generation has underscored the significance of multi-modal video generation approaches. Drawing inspiration from the multi-sensory nature of human learning, where the integration of different senses enhances comprehension, multi-modal video generation incorporates diverse data types—such as text, audio, images, video, depth, and body pose—to achieve finer control over various aspects of video creation. However, the effective fusion of these modalities continues to pose significant challenges.

Training large generative models is highly resource-intensive and the computational and memory requirements escalate dramatically as the complexity and dimension of the data increase, making it challenging to scale models effectively [9]. For instance, training Llama 3-70B LLM model, required 6.4 million H100 GPU hours [10]. Similarly, given the demand for high-resolution, long-duration videos, training video models could require hundreds of thousands to millions of GPU hours, depending on the model architecture and scale. For example, to train the Open-Sora1.2 [11]- a model with 1.1B parameters generating 97-frames videos- 4.8k Ascend and 37.8k H100 GPU hours was required [12].

*Vikit.ai
†University of Paris-Est
Corresponding author e-mail: elnaz@vikit.ai

To alleviate these challenges, a promising line of research focuses on developing techniques that can be integrated into existing open-source models to enhance their performance with minimal additional training [13]. Moreover, researchers are working on optimizing the training as well as inference process by introducing more efficient architectural designs, refined training strategies, quantization techniques, and better-curated datasets, all aimed at reducing computational costs while maintaining or improving model performance.

This survey presents a comprehensive review of contemporary research in video generation, providing a structured overview of the field's latest advancements to foster its future development. As illustrated in Fig. 1, the rest of the paper is structured as follows. Chapter II offers a review of video generation fundamentals. Chapter III delves into state-of-the-art models for video generation in three main topics: video synthesis, video editing, and enhancement techniques. Chapter IV and V discuss evaluation metrics and datasets pertinent to video generation, respectively. Finally, chapter VI addresses key challenges and outlines potential directions for future research in this rapidly evolving domain.

II. Video generation fundamentals

A. Generation strategies

From a problem-solving perspective, the video generation task can be tackled through various strategies.

1) Divide & conquer (Cascade): In this strategy, the model begins by generating distant keyframes that outline the overall video storyline of the video. The gaps between keyframes will be later filled using Spatio-Temporal super-resolution modules. A key advantage of this scheme is its memory efficiency as the modules in the cascade can be trained independently and in parallel. However, maintaining global temporal coherence remains challenging, especially in cases of fast motion which can lead to temporal aliasing. Additionally, the super-resolution modules may suffer from a domain gap as they are trained on real frames but applied to synthetic frames during inference [14]. Notable models using this approach include LaVie [15], Nuwa-XL [16], and ImagenVideo [17].

2) Brute-force: This approach processes all frames simultaneously, often at a lower spatial resolution, to ensure globally coherent motion across the entire video. Spatial Super Resolution (SSR) models are then applied to generate high-resolution outputs [14].

3) Auto-regressive: This strategy focuses on directly generating detailed frames conditioned on preceding ones. Examples of model following this approach include CogVideo [18], Nuwa-Infinity [19], MCVD [20].

4) Hybrid: Each of these strategies has distinct strengths and weaknesses. Autoregressive approaches excel at capturing long-range dependencies and ensuring coherent motion. However, as the videos are generated sequentially, they can be computationally expensive and prone to quality degradation due to cumulative errors. Divide & Conquer strategy offers greater efficiency and scalability through parallelization but struggles to maintain coherence across video segments for longer video generation. Some models like Nuwa-XL [16] adopt a hybrid approach by generating distant key-frames and filling in the gap between keyframes autoregressively.

B. Core frameworks

1) Diffusion Models: Once a strategy is selected, various machine learning **frameworks** can be utilized to implement it effectively. Fig. 3, illustrates some of the most widely used GenAI frameworks in the video generation, including Diffusion models, Generative Adversarial Networks(GAN) [21], Variational Auto Encoders (VAEs), and Flow-Based Generative models.

Preliminary: Diffusion models are the most dominant choice among existing video generation frameworks. One of the most popular approaches for implementing diffusion models is Denoising Diffusion Probabilistic Models (DDPM) [6] that encompasses a forward and a backward process. During training, the forward diffusion process involves iteratively adding sampled noise to the initial input x_0 over T steps, following a Markov chain. The Markov property guarantees that the noisy frame x_t at time step t only depends on the frame x_{t-1}:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

where I is the identity matrix and β_t is calculated by a variance-preserving noise scheduler, to gradually intensify the noise. Ho et al [6] proposed a formula to directly calculate the distribution at a given step t:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha_t}}x_0, (1-\bar{\alpha_t})I) \tag{2}$$

Whereas $\bar{\alpha_t} = \prod_{s=1}^{t} \alpha_s$ and $\alpha_t = (1-\beta_t)$.

In the backward process (inference time), the objective is to iteratively denoise the noise image x_T back to a clear image x_0. We can approximate $q(x_{t-1}|x_t)$ by a parametrized model p_θ:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3}$$

The reverse process model is trained with the variation lower bound on negative log-likelihood of x_0 which involves the KL divergence between q and p_θ. Putting in place a few simplifications (such as $\Sigma_\theta = 1$) [6], μ_θ can be approximated by a **denoising neural network** ϵ_θ. Thus the training loss would be simplified to a mean-squared error between the predicted noise $\epsilon_\theta(x_t)$ and the ground truth sampled Gaussian noiseϵ:

$$L_t^{simple} = \mathbb{E}_{x_0,t,\epsilon}[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha_t}}x_0 + \sqrt{1-\bar{\alpha_t}}\epsilon, t)\|^2] \tag{4}$$

Backbone architecture: There are multiple variations of diffusion models, each differing based on the architecture chosen for the denoising neural network. A common approach among researchers is to use a *convolutional U-Net* architecture as the default backbone. Some works such as Make-a-Video [22] opted for 3D or pseudo-3D convolutions to incorporate temporality into these architectures. However, *Vision Transformers (ViTs)* [23], have gained prominence in models like Sora [24] and Latte [25], due to their superior scalability and performance in computer vision tasks compared to traditional convolutional networks [26]. A ViT takes as an input the image

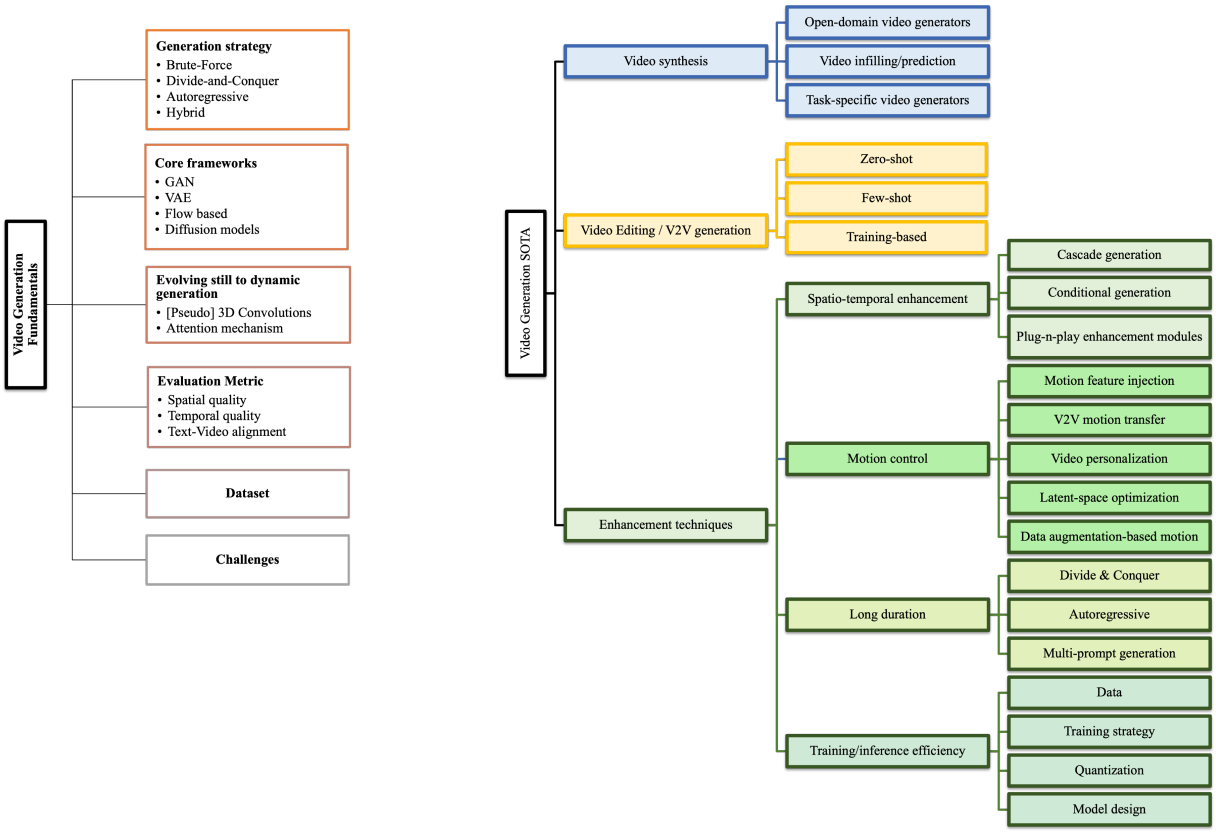

Fig. 1. A taxonomic overview of fundamental concepts and state-of-the-art in video generation from the survey

$I \in \mathbb{R}^{H \times W \times C}$ that is reshaped into a sequence of 2D patches $I_p \in \mathbb{R}^{N \times P \times P \times C}$ where W, H and C represent the spatial dimensions, and number of channels respectively. (P, P) is the resolution of each image patch and $N = HW/P^2$ which are then flattened and linearly embedded into tokens. *Diffusion Transformers (DiTs)* [26] explores incorporating ViTs into the standard diffusion model pipeline to improve performance as an alternative to U-Net.

Tokenizer: When using vision transformers, one crucial component is the visual tokenizer which maps pixel-space inputs into discrete tokens suitable for the transformers. This step is critical for enabling the attention mechanism (discussed in section II-C) of transformers to effectively process visual data. The VQ-VAE (Vector Quantized Variational Autoencoder) is a pivotal work in this regard which employs vector quantization to learn discrete representation of image data. It includes a fixed codebook of vectors, where the encoder's output is matched with the closest vector based on Euclidean distance. [27]. Tokenizing video data presents greater challenges compared to images. Models such as MAGVIT [28] and MAGVIT-v2 [29] have aimed to extend and enhance VQ-VAE to address these challenges.

Pixel spaces vs latent space: It's important to note that for diffusion models operating directly in pixel space, both training optimization and inference process can become resource-intensive because of the need for sequential evaluations [30]. Latent Diffusion Models [30] address this issue by first learning an auto-encoder to compress images into a lower-dimensional latent space. With the auto-encoder frozen, the diffusion model is trained on this latent representation, significantly reducing the computational burden.

2) Generative Adversarial Networks: **GANs** consist of two main components; a generator that generates synthetic data given a random noise, and a discriminator that distinguishes between synthetic and real samples. The objective is for the generator, once training is over, to produce highly realistic samples. The overall loss function for the framework is formulated as:

$$\max_D \min_G (\mathbb{E}_x[logD(x)] + \mathbb{E}_z[log(1 - D(G(z))]) \quad (5)$$

where x represent real data and z is a noise sampled from a prior distribution $p_z(z)$. While the video synthesis capabilities of GANs have been investigated in works such as MoCoGAN [31] and StyleGAN-V [21], these methods face challenges when dealing with complex scenes and multiple objects.

3) Variational Auto Encoders: **VAEs**, utilize an encoder $\mathcal{E}$ to map the input data into a distribution in the latent space, which the decoder then uses to sample from and reconstruct the input:

$$\mathcal{E} : q(z|x), \mathcal{D} : p(x|z) \quad (6)$$

where maps $\mathcal{E}$ the input data x to a latent space distribution, and $\mathcal{D}$ generates data from a point sampled from the latent

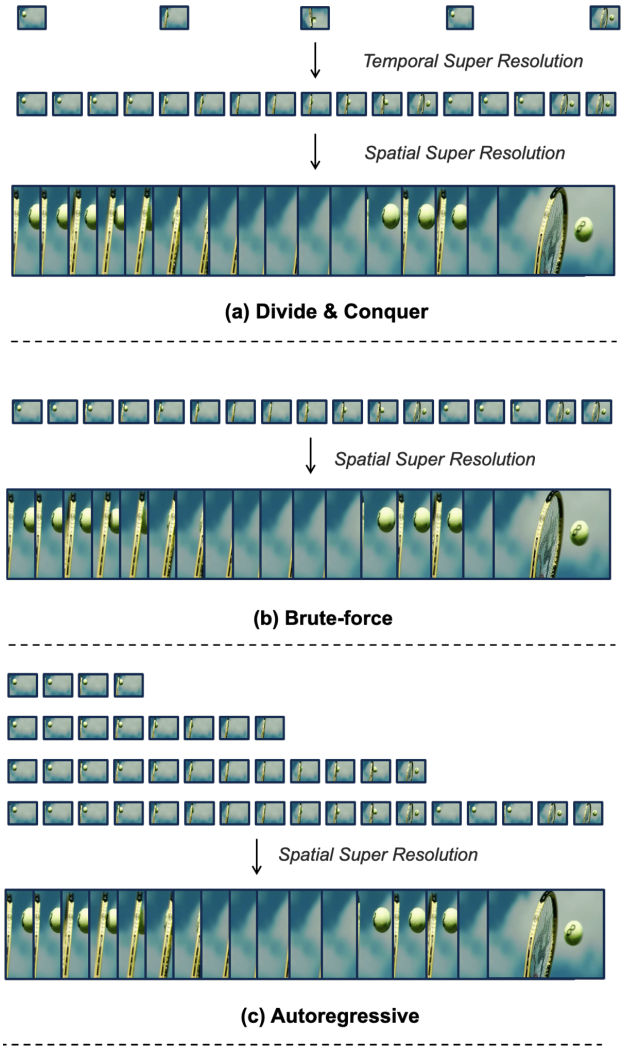

Fig. 2. Video generation strategies overview

distribution. The learning objective function for VAE can be formulated as follows:

$$L_{VAE}(x,z) = L_{reconstruction} + D_{KL}(q(z|x), p(z)) \quad (7)$$

$$L_{reconstruction} = \mathbb{E}_{q(z|x)}[\log p(x|z)] \quad (8)$$

where D_{KL} measures how much the latent distribution $q(z|x)$ approximated by the encoder, deviates from the prior distribution $p(z)$ which is typically a standard normal distribution. In practice, this often reduces to the MSE between the original data x and the reconstructed data. VAEs can be divided into two groups, discrete and continuous latent based on the types of the quantization. Continuous VAEs have no quantization, while discrete VAEs learn a codebook for quantization and use it to convert the continuous latent features to discrete indices, called VQ-VAE.In video generation, 2D VAEs are often extended into 3D by incorporating 3D convolution or temporal attention mechanisms, as discussed in section II-C.

4) Flow-Based Generative Models: Flow-based generative models employ a series of invertible transformations to model the data distribution directly. They offer exact log-likelihoods,

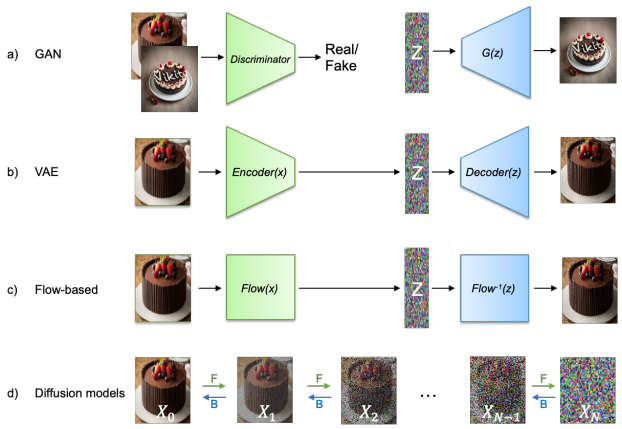

Fig. 3. An overview of most common generative AI core technologies

contrasting with approaches such as GANs and VAEs, which rely on approximate likelihoods. Given input data x, a flow-based model aims to map it to a latent variable z through a composition of invertible functions $f = f_1 \circ f_2 \circ \cdots \circ f_K$, yielding $z = f(x)$. This invertibility allows us to express the likelihood $p(x)$ through a change of variables rule:

$$\log p(x) = \log p_\theta(z) + \log \left| \det \left(\frac{\partial f}{\partial x} \right) \right| \quad (9)$$

where $p_\theta(z)$ is a prior distribution, often Gaussian, defined over z, and the term $\frac{\partial f}{\partial x}$ is the Jacobian matrix of the transformation f with respect to x. The determinant of this Jacobian plays a crucial role as it adjusts the density under the transformation. However, calculating the Jacobian determinant becomes computationally intensive for high-dimensional data, especially in video generation tasks, due to the need for precise frame-by-frame consistency.

Continuous Normalizing Flows (CNFs): Continuous Normalizing Flows (CNFs) [32] offer an innovative approach by modeling the transformation f as a continuous flow. Rather than applying discrete transformations, CNFs parameterize the data transformation as a time-dependent process using an ordinary differential equation (ODE):

$$\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x)) \quad (10)$$

where $t \in [0,1]$ and $\phi_0(x) = x$ represents the initial condition. The vector field v_t then can be modeled with a neural network $v(x,t;\theta)$ that acts as a time-varying vector field. In essence, $v(x,t;\theta)$ defines the instantaneous direction and speed at which $\phi_t(x)$ should evolve. By treating transformation as a continuous function, CNFs allow for more flexible modeling of complex distributions. Training CNFs involves solving this ODE, which requires expensive numerical integration. This integration step, typically performed by an ODE solver, makes it challenging to scale CNFs to high-resolution or high-dimensional data, such as video frames.

Flow Matching (FM): Flow Matching (FM) [33] is an alternative approach designed to simplify the training of CNFs by circumventing the need for direct ODE simulation. Instead of solving the ODE explicitly, FM defines a target probability path from a simple initial distribution (e.g., standard Gaussian)

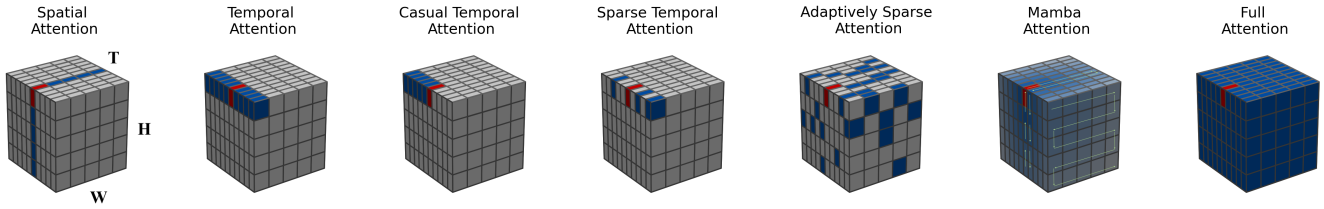

Fig. 4. Attention mechanism comparison- The red patch represents the query patch, while the blue tokens indicate those interacting with the query. For simplicity, only the unidirectional scan is depicted in Mamba attention and darker blue shades represent more intense interactions with the query patch.

to a complex target distribution. The objective function of FM is to align the model's vector field $v_t(x)$ with the target vector field $u_t(x)$ that describes the desired evolution of the probability density over time. The training objective is:

$$L_{FM}(\theta) = \mathbb{E}_{t,p_t(x)}\|v_t(x) - u_t(x)\|^2 \tag{11}$$

where $u_t(x)$ represents the target vector field that enables the probability density path $p_t(x)$ to transform smoothly from a simple base distribution p_0 (e.g., $\mathcal{N}(0,1)$) to a complex distribution approximating the data. A challenge here is that the closed forms of $p_t(x)$ and $u_t(x)$ are typically unknown. To address this, the Conditional Flow Matching (CFM) [33] approach provides a simple simulation-free training objective:

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t,q(x_1),p_t(x_t|x_1)}\|v_t(x_t) - u_t(x_t|x_1)\|^2 \tag{12}$$

where $u_t(x_t|x_1)$ represents a time-variant conditional probability path p_t toward target data sample x_1. Similar to diffusion models, training samples can be generated by sampling data points from a known distribution and linear interpolation between data points and noise:

$$x_t = tx_1 + (1-t)x_0 \tag{13}$$

where $x_1 \sim q$ is a sample from the target data distribution and $x_0 \sim \mathcal{N}(0, I)$. The CFM loss regresses the model vector field $v_t(x;\theta)$ onto $u_t(x)$, focusing on matching flows rather than directly learning probability density functions.

Applications in Video Generation: Flow-based models, particularly in conjunction with Flow Matching techniques, demonstrate considerable potential in video generation by modeling the complex distributions of high-dimensional video data [12], [34], [35] . The invertible nature of these models supports precise control over transformations, offering a promising direction for generating temporally consistent frames. Nevertheless, computational limitations persist, as flow-based methods require high memory and computational resources, especially for long-duration or high-resolution video sequences. The development of efficient flow-based techniques, such as Flow Matching, provides a path forward for enhancing video generation capabilities in multimodal and dynamic contexts.

C. From Stills to Motion: Advancing Image Models for Video Generation

AI models achieve high performance in T2I generation taking as input batches of 3D tensors of images of shape $I \in \mathbb{R}^{b\times h\times w\times c}$. However, video generation tasks are coupled with treating batches of 4-dimensional video tensors $V \in \mathbb{R}^{b\times f\times h\times w\times c}$. Several approaches to incorporating the temporal dimension into models include employing 3D and pseudo-3D convolution, factorization, temporal attention, and positional embedding [36]. Training architectures containing 3D convolutions are often considered computationally expensive. On the other hand, the pre-trained image layers in the T2I models capture high-quality content priors that are desirable to be leveraged into T2V generation task. Therefore, one preferable way to bring the temporal aspect into the game is to inflate T2I networks and let these spatial layers deal with batches of video frames independently. For example in AnimateDiff [37], similar to [38], [39], for training the temporal modules data is reshaped into $b.h.w \times c \times f$, whereas for training spatial modules it is reshaped into $b.f \times c \times h \times w$ so that T2I spatial layers interpret the video as a batch of independent images. Video Latent Diffusion Model(V-LDM) [38] proposes a slightly different approach in which the temporal modules consist of temporal attention as well as residual blocks based on 3D convolutions that can process the output of the spatial modules in video format $b \times c' \times f \times h' \times w'$ (where c', h', w' represent dimensions of spatial feature space).

As for attention mechanisms, several ideas have been explored in the literature. Fig. 4 illustrates some of the popular attention mechanisms. Factorized attention is a common approach that separates spatial and temporal attention to reduce computational complexity and facilitate fine-tuning, a.k.a Axial Attention [40]. Instead of computing attention over the entire 2D or 3D space at once, axial attention computes attention along one axis at a time (1D). In the context of image/video processing, since the length of any single axis -that is, the height or width of an image- is typically much smaller than the total number of elements, an axial attention operation brings in a significant saving in computation and memory over standard self-attention. For instance, authors in [22], decouples the spatial and temporal dimensions and processes video data using a combination of 2D spatial attention (which operates on each frame individually) followed by 1D temporal attention (which captures dependencies between frames).

***Full Attention**:* Separating spatial and temporal attention requires extensive implicit transmission of visual information, significantly increasing the learning complexity and making it challenging to maintain the consistency of large-movement objects. To address these issues and inspired by long-context training in LLMs, authors in [41] propose **3D Full Attention**,

a 3D text-video hybrid attention mechanism that can be accelerated by parallelization techniques.

Sparse Attention: As the full attention mechanism remains heavy in computation, some researchers have tried to come up with a hybrid approach that maintains the 3D attention span to some extent. Sparse attention is a family of attention mechanisms, that aims to reduce the number of tokens processed during attention to improve computational efficiency. For instance, 3D Nearby Attention [42] focuses on local regions within a three-dimensional space. This mechanism enhances the model's ability to capture fine-grained spatial details by attending primarily to nearby tokens or features. This localized approach helps reduce computational complexity while maintaining relevant contextual information. Three-dimensional sparse attention from Godiva [43] is another example, where each token in the input sequence can attend to a limited number of other tokens based on predefined strategies allowing a mix of local and controlled global attention.

Mamba Attention: Mamba [44] is a State-Space Model, that has recently gained prominence in deep learning for its universal approximation capabilities and efficient long-sequence modeling. That makes it particularly interesting for the video domain as it inherently involves long temporal sequences. ZigMa [45] introduces a simple zigzag scanning method that rearranges the scan path of Mamba in a heuristic manner to improve spatial and temporal continuity. Authors of Matten [46] demonstrated the effectiveness of combining Mamba and temporal attention to capture global and local temporal relationships respectively.

Cross Attention: is also widely used in video generation models to facilitate interaction between different modalities or sequences by enabling tokens from one sequence to attend to another. This enhances multi-modal learning by allowing the model to build richer associations between diverse inputs. For example, in the joint image and text-to-video generation, cross-attention enables the model to align relevant visual features from the image with corresponding linguistic elements from the text, driving a more coherent and context-aware generative process, which leads to higher-quality and more accurate outputs [47].

III. State-of-the-art

We categorized state-of-the-art video generation methods into three primary categories: video synthesis, video editing, and enhancement techniques. The video synthesis section covers models that generate videos from alternative modalities, while the video editing section focuses on models designed for video-to-video generation. The enhancement section highlights techniques and models to address specific limitations in existing video generation approaches. Table II summarizes the characteristics of the mainstream video generation models.

The video generation task can be formalized as follows. Given an input X and a set of conditions $\mathbb{C}$, we aim to generate a video $V \in \mathbb{R}^{T\times C\times H\times W}$, where T denotes the number of frames, W and H represent the spatial dimensions of each frame, and C is the number of channels, typically 3 for RGB videos. We define a model $\mathbb{F}_\theta$ parameterized by θ that maps the input and conditions to the target video:

$$\mathbb{F}_\theta : (X, \mathbb{C}) \rightarrow V \tag{14}$$

To train the model, the objective is to minimize a loss function $\mathcal{L}(\mathbb{F}_\theta(X, \mathbb{C}), V)$ that quantifies the difference between the generated and target video. Inputs X and conditions $\mathbb{C}$ may include modalities such as text, images, video sequences, pose information, or depth maps. The optimization goal is then:

$$\theta^* = \arg\min_\theta \mathcal{L}(\mathbb{F}_\theta(X, \mathbb{C}), V) \tag{15}$$

The goal is to find optimal parameters θ such that $\mathbb{F}_\theta$ generates V by effectively interpreting X and $\mathbb{C}$ as guiding inputs.

A. Video synthesis

Video synthesis involves generating new videos from various input sources- such as text and images- ideally producing coherent and dynamic visual content. The problem formulation follows Eq. 15 where X is a text sequence $X \in \mathbb{R}^d$ for text-to-video and an image $X \in \mathbb{R}^{C\times H\times W}$ for image-to-video generation schema. In this section, we review various video synthesis approaches including general-purpose and task-specific video generation and video infilling models.

1) Mainstream open-domain video generation models: **Text2Video (Text-conditional video generation)**: One of the primary challenges in training video generation models is the scarcity of large high-quality paired video-text datasets. In contrast, vast and diverse datasets [48] are readily available for T2I generation tasks. Several studies have attempted to overcome this limitation by leveraging the existing highly efficient models in the T2I domain. Nevertheless, these models inherently lack the temporal aspect required for video generation. To address this issue, researchers have adapted the architecture of T2V models by different means such as 3D or pseudo 3D convolutions or temporal attention mechanisms(section II-C). For example, *CogVideo* [18] uses a *frozen* pre-trained T2I model for autoregressive transformer-based T2V generation. Incorporating temporal layers into the frozen spatial layers, reduces memory usage during training, requiring only a few trainable parameters.

Make-a-Video [22] on the other hand, fine-tunes the T2I base model and establishes a joint text-image prior, eliminating the need for paired text-video data. To enhance the temporal consistency of the generated videos, the authors propose a spatio-temporal factorized diffusion-based architecture, using pseudo-3D convolution and temporal attention layers. These temporal layers are fine-tuned on unlabeled video data, while the other modules are trained on image data alone.

Imagen video [17] employs a cascading architecture that enables the progressive generation of higher-resolution, longer videos. The base video diffusion model is conditioned on text embeddings from a large frozen language model. To ensure temporal consistency with lower memory and computational demands, temporal convolutions are applied in the super-resolution modules, while a temporal attention mechanism is implemented in the base diffusion model to maintain global coherence. The model is jointly trained on both image and

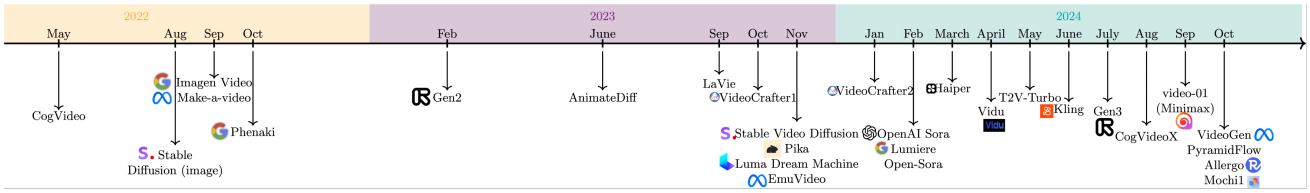

Fig. 5. Timeline of release of mainstream video generation models

video datasets, with individual images treated as single-frame videos.

Phenaki [49] introduces C-ViViT, an autoregressive transformer-based autoencoder that compresses variable-length videos into tokens using a causal attention mechanism. To generate video tokens from text, it utilizes a bidirectional masked transformer conditioned on pre-computed text tokens produced by a frozen pre-trained text encoder. These video tokens are at the end de-tokenized to create the actual video. The model progressively fills in tokens through an iterative process, starting with all tokens masked and predicting a subset of them at each step. By incorporating a new text prompt along with the last few frames of the preceding video segment, the model can extend the video's duration. For training, Phenaki leverages a large corpus of image-text pairs, supplemented with a smaller set of video-text data. To further enhance spatio-temporal resolution, temporal interpolation and video super-resolution techniques are applied.

Text2Video-zero [50] introduces a training-free method to animate a pre-trained T2I model through latent wrapping based on a predefined affine matrix. This approach enriches the latent codes of generated frames with motion dynamics to ensure temporal consistency. Additionally, frame-level self-attention is reprogrammed by applying a novel cross-frame attention mechanism, where each frame attends to the first frame, thereby preserving the overall context.

AnimateDiff [37] proposes a *plug-and-play* motion module that can be integrated into any personalized T2I model without model-specific fine-tuning. The training of AnimateDiff involves 3 components: a domain adapter, a motion prior, and an optional MotionLoRA module. To avoid learning the quality discrepancies between T2I and T2V datasets in the motion module, the authors propose to fit the domain information to a domain adapter network, which can be partially or integrally discarded during inference, mitigating any negative effects. To model motion dynamics, the approach inflates the T2I model as detailed in section II, and introduces a temporal transformer module with several self-attention blocks along the temporal axis. Finally, MotionLoRA, a lightweight fine-tuning technique, is optionally proposed to adapt pre-trained motion modules to new patterns such as camera zooming, panning, and rolling. This adaptation requires as few as 20-50 reference videos, 2000 training iterations (approximately 1-2 hours), and minimal storage space ($\sim$30MB). The integration of existing controllable generation methods, such as ControlNet, is also possible due to the decoupling of visual content and motion priors. However, the model's generated motions are often constrained by the motions present in the training data, leading to generic motion patterns that may not always align with the provided text prompts.

LaVie [15] is a cascade Latent Diffusion Model (LDM) built on a pre-trained T2I foundation. It captures temporal information through temporal self-attentions with rotary positional encoding and pseudo-3D convolution in its LDM. Fine-tuned jointly on images and videos, LaVie also includes an LDM upsampler, utilizing a diffusion-based image x4 more upscale as prior, to boost video resolution up to $1280x2048$ pixels.

EmuVideo [51] adapts a T2I U-Net prior augmented with temporal parameters (Pseudo-3D conv and temporal attention). Identity initialization of these parameters improved the model convergence by a factor of two. The model employs two frozen text encoders (CLIP and 5-XL) to encode text prompts. During training on video-text pairs, the model samples an initial frame, I, and predicts subsequent frames based on the text prompt and the conditioning image I. This approach allows EmuVideo to generate videos conditioned on given text and image prompts. Fine-tuning on a small set of high-motion videos proved to further enhance the motion quality in the generated videos.

Stable Video Diffusion [52] is a Latent Video Diffusion model designed for T2V / I2V generation, with adaptability to camera motion-specific LoRA modules. The authors introduce a three-stage training process encompassing text-to-image pre-training, video pre-training, and high-quality video fine-tuning. Underscoring the critical role of data curation, they present a systematic workflow comprising several key steps. A cut detection pipeline is employed to remove cuts and transitions, resulting in x4 more clips. Each clip is then annotated using three synthetic captioning methods: middle-frame image captioning, video-based captioning, and an LLM-based summarization of the first two captions. Clips with excessive on-screen text or motionless ones are further filtered out using optical character recognition and optical flow score, respectively.

Videocrafter1 [53] is built upon a T2I model by incorporating temporal attention layers into the SD UNet architecture to capture temporal consistency. It proposes a full patch visual tokens conditioning using CLIP last layers that represent more details, to maintain higher fidelity to the conditioning image. Its joint image-video training schema includes low resolution pertaining, then progressively fine-tuning on higher resolutions. Image and text conditioning inputs are leveraged via a cross-attention mechanism. *Videocrafter2* [54] is an improved version designed to address the limitations posed by high-quality data scarcity. Unlike previous approaches that begin with high-quality image-based pretraining, Videocrafter2 first

trains a video model using a large volume of low-quality videos, followed by fine-tuning with high-quality images. Experimental results indicate that the strong spatial-temporal coupling in the video model, achieved by fully training both spatial and temporal layers, allows the model to tolerate parameter perturbations in both layers without significant motion degradation.

ModelScopeT2V [55] initializes the spatial part from the Stable Diffusion model, and proposes the factorized spatio-temporal blocks that empower the capacity of temporal dependencies. The core of this model is an LDM using VQGAN latent space and cross-attention blocks for text conditioning using CLIP embedding of the text prompt.

CogVideo [18] builds on the pre-trained T2I model, CogView2, by introducing a multi-frame-rate hierarchical training strategy aimed at better-aligning text with video clips. This approach involves two key components: a sequential generation model and a frame interpolation model. The sequential model generates keyframes based on the input text, while the interpolation model recursively fills in the intermediate frames by adjusting the frame rates, ensuring smooth and coherent video sequences. Additionally, the incorporation of the Swin attention mechanism improves parallel generation across distant regions of different frames, significantly accelerating the auto-regressive generation process. *CogVideoX* [41] pushes further the quality in terms of visual context, motion, and duration. It is a large-scale diffusion transformer that uses a 3D Variational Autoencoder (VAE) architecture to compress videos across both spatial and temporal dimensions, and expert transformers to facilitate the deep fusion between the two modalities. To further improve the transformers' performance, CogVideo extends RoPE encoding to video, by applying 1D-RoPE independently to each dimension of the video latent (x, y, t) and then concatenating the results along the channel dimension to create the final 3D-RoPE encoding. Its progressive training techniques, which involve initial training on short videos followed by fine-tuning on longer ones, enable the model to generate extended video sequences. The authors introduced a pipeline for dense video caption data generation to enrich training data.

Unlike the common divide & conquer cascade architecture typically used for T2V generation, *Lumiere* [14] processes all frames simultaneously, without relying on a cascade of Temporal Super-Resolution models, to ensure globally coherent motion across the entire video. Then to achieve high-resolution output, a Spatial Super Resolution(SSR) model is applied on overlapping windows. Lumiere's Temporal UNet architecture is built upon a pre-trained T2I model, which is kept fixed during training, by interleaving temporal blocks (temporal conv + temporal attention) and up/down scale modules. Lumiere is also capable of image-conditional video generation and stylization.

More recently, the Flow Matching paradigm has demonstrated superior performance and faster generation speeds compared to traditional diffusion models, along with increased robustness to noise schedule choices. One notable example is Video Gen [35], a 30B parameter transformer model trained with a maximum context length of 73K video tokens using flow-matching paradigm. Meta Video Gen follows multi-stage training including joint image-video training, followed by supervised fine-tuning on a curated set of high-quality text-video pairs. This approach allows it to generate personalized videos, including those conditioned on an individual's face through additional post-training. Additionally, Video Gen incorporates a camera motion classifier to predict 16 distinct types of camera motion. Pyramid Flow [12] is another example, introducing the pyramidal flow matching algorithm, which breaks down the video generation process into multiple stages where only the final stage operates at full resolution, resulting in significantly reducing computational demands.

In addition to research advancement in open-source, several commercial models - such as Runway Gen3 [47], [56], Pika [57], Haiper [58], Kling [59], Vidu [60], Sora [24], Veo [61]- and Minimax [62] have emerged, offering diverse tools for high-quality video creation and editing. While these models bring powerful video generation capabilities, details about their underlying technologies are often not disclosed.

Image-to-Video generation: I2V represents another major paradigm in video synthesis, where the main challenge is to preserve the identity of the entities in the reference input image while producing high motion dynamic with high temporal consistency. Researchers have explored I2V generation through various methods. One approach is to use physical simulation, which despite its accuracy lacks generalisability. Some other methods animate still images by leveraging explicit or implicit image-based rendering, often guided by estimated motion fields or geometric priors. Alternatively, some methods predict future video frames from single images by learning spatiotemporal priors - either from a T2V model or through temporal guidance from a conditional signal, such as video or pose. Novel View Synthesis methods could also be used for I2V generation by synthesizing different viewpoints from given source images, following a specified camera pose trajectory [63]. Yet, many of these techniques are constrained by their focus on specific types of motion or objects.

I2VGen-XL [64] is an I2V model that employs a cascade of 2 LDMs. In the base stage and at lower resolution, two hierarchical encoders are employed to simultaneously capture high-level semantics and low-level details of input images. In the refinement stage, a separate LDM is utilized to enhance the resolution and temporal continuity of videos. This model however faces challenges in terms of runtime as well as preserving visual details of the input image, due to insufficient context understanding and loss of information of the input image. To address this, *DynamiCrafter* [65], introduces a diffusion model-based I2V framework, builds on the T2V model VideoCrafter [53], to animate a still image using T2V generative priors. It integrates the input image into the generative process as guidance, using a dual-stream image injection mechanism via cross-attention. This mechanism is designed to retain visual details and process the input image in a context-aware manner. To further enhance the model's capacity to incorporate image conditions across different layers, learnable coefficients are introduced to fuse text and image-conditioned features. This design is based on the observation that intermediate U-Net layers are more associated with object

shapes and poses, while the outermost layers are more closely linked to appearance. Moreover, guiding the denoising U-Net with the conditional image and per-frame initial noise has been shown to significantly improve visual coherence. PIA [66] is a Plug-n-Play image animator that excels in aligning with condition images built upon a base T2I model. It achieves motion controllability by text using its temporal alignment blocks.

Building I2V models upon T2V models often involves full or partial training of the core weights in a large T2V model which is both costly and restrictive. To mitigate this issue, *I2V-Adapter* [67] proposes a lightweight plug-n-play adapter that transfers the input image to subsequent noisy frames via a cross-frame attention mechanism, preserving the identity of the input image without modifying the pre-trained T2V model. This approach is conceptually similar to IP-Adapter [68] in T2I. Additionally, the I2V-Adapter incorporates a Frame Similarity Prior, which regulates the balance between motion amplitude and video stability through two adjustable control coefficients.

Another challenge is to effectively integrate image features into the T2V base model. A common approach is to use the CLIP visual encoder to extract semantic features from images. However, this method often compromises the preservation of content and structure from the input image in the generated videos. Since CLIP is primarily trained to align visual and linguistic features, it prioritizes high-level semantics, potentially at the cost of fine-grained details. To address this limitation, some methods have introduced supplementary features, such as incorporating full visual tokens from the last layer of the CLIP image encoder or utilizing learnable features to retain the intricate details of the original images more effectively [64], [65].

Multi-modal video generation: Although diffusion-based approaches are dominant for video synthesis tasks, recently some frameworks have emerged exploring the potential of Large Language Models (LLMs), to generalize the task of X-to-X or multimodal-conditioned generation and address the appetite for a multi-usage foundation model in video/image generation. The advantage of this approach is that LLMs can flexibly incorporate numerous tasks. However, the challenge is that as the model size grows, training data must grow. Lumina-T2X is a Flow-based Large Diffusion Transformers [69] proposing a unified framework for T2X (e.g., text-to-image, text-to-video, text-to-audio) generation, targeting to be a chatGPT for vision. It tokenizes any modality regardless of resolution, aspect ratio, or even temporal duration into a unified 1-D token and then processes the one-dimensional sequences, similar to the way LLMs process natural language. VideoPoet [70] is a model for synthesizing videos from a variety of conditioning signals such as text, image, depth, and optical flow. Its autoregressive Transformer-based framework is trained in two stages of pretraining and task-specific adaptation, similar to LLMs. It employs the MAGVIT-v2 [29] tokenizer for joint image and video tokenization to reduce the sequence length required by the LLM. The SoundStream tokenizer is used for audio and the text modality is embedded by a T5-XL text encoder. After converting the image, video, and audio modalities into discrete tokens within a shared vocabulary, a language model with a decoder-only architecture is directly used to generate videos and audio, in the token space. A non-autoregressive video transformer super-resolution is then applied in token space, as the sequence length can get too long to be processed in an autoregressive manner.

2) Video Infilling / Prediction: : The problem of video synthesis is viewed in the form of video infilling or frame interpolation [71], where the objective is to predict the missing frames between the given start and end frames. To formalize this more precisely, Eq. 14 should be reformulated as follows:

$$\mathbb{F}_\theta : (I_s, I_e, \mathbb{C}) \rightarrow V \tag{16}$$

Where $V = \{I_s, I_1, I_2..., I_N, I_e\}$. This can also be seen as a way to augment the quality of generated videos by complementing a video generation model that does not have a high enough frame rate. Similarly, video prediction(completion) could be defined as:

$$\mathbb{F}_\theta : (I_s, \mathbb{C}) \rightarrow V \tag{17}$$

Where $V = \{I_s, I_1, I_2..., I_N\}$. Recently, several diffusion-based frame interpolation models, building upon T2V and I2V frameworks, have emerged. For instance, SEINE [72], based on the pre-trained T2V diffusion model LaVie [15], recursively uses the last few frames of a generated video to predict subsequent ones, ensuring semantic consistency with the initial frame while preserving temporal coherence and text alignment. However, this approach can produce uncanny morphing effects, especially with humans, when consecutive frames lack strong similarity. ToonCrafter [73], built on DynamiCrafter, is a diffusion model tailored for cartoon interpolation. Similarly, I2V models like DynamiCrafter [65], SparseCtrl [74], and PixelDance [75] show adaptability for video interpolation/transition tasks, by concatenating two input frames with noisy frame latent [65], [75] or using an auxiliary frame encoder [75]. MCVD [20] uses random frame masking, making it capable of handling a range of video generative modeling tasks, including video prediction and interpolation. However, the output videos can still become blurry or inconsistent when the number of generated frames is very large. FILM [71] is a prominent frame interpolation algorithm, especially adept at managing large motion. The algorithm initiates by predicting bidirectional optical flow and context maps, which serve to capture both motion dynamics and occlusions between consecutive frames. Following this, a multi-scale warping mechanism is employed to align features across various resolutions, thereby effectively accommodating large displacements. Finally, a refinement network is applied to enhance the quality of intermediate frames, significantly reducing artifacts. Vidim is a Video interpolation [76] with a two-stage cascade diffusion, where it first generates a low-resolution video and subsequently refines it to high resolution, enhancing detail and fidelity in the output.

3) Down-stream task-specific video generation models: Similar to the recent trend toward the development of smaller, specialized large language models (LLMs), video generation models are increasingly being tailored for specific downstream

tasks. These models are designed to perform specialized functions, such as panorama text-to-video generation [77], time-lapse video generation [78] or *word visualizer* [79]. Another prominent application is character or face animation, driven by various input signals such as video, audio, images, text, or pose sequences.

Character animation models are particularly focused on generating videos from static images, with the driving signals guiding the motion. These characters may represent real humans, animated figures [80]–[83], or animals [84]. Depending on the nature of the target motion, some models specialize in generating dance sequences [85]–[87] while others are trained to create fashion pose videos [88]. Face and portrait animation models, on the other hand, face distinct challenges, such as accurately capturing the nuances of human facial expressions, preserving the unique attributes of individual faces, and maintaining identity fidelity, and accurate lip synchronization with driving audio, video, or input text [89]–[100]. Training models for downstream tasks often necessitates the use of task-specific labeled datasets. To mitigate this requirement, reward-based fine-tuning techniques have been developed to adapt foundational video generation models [101], [102]. For instance, VDA [102] utilizes reward gradients derived from pre-trained discriminative models to refine video diffusion processes. This approach enhances the alignment of generated videos with specified inputs, such as text or images, while streamlining the training procedure by minimizing reliance on large-scale labeled datasets.

In summary, several innovative approaches have emerged to address the challenges of video synthesis. Techniques such as frame rate conditioning [18], [22], [53] and factorization [18], [22], [51], [52] are widely utilized to adapt T2I models for video generation. Another significant approach involves using expert models to produce visual content at a coarse frame rate, followed by interpolation techniques to enhance this to smoother, high-frame-rate outputs.

Training strategies for integrating T2I models into T2V synthesis often follow one of two approaches: partial training, where only temporal modules are fine-tuned while spatial modules remain fixed; and full training, in which both spatial and temporal components are trained together, using image-model weights as initialization. However, a significant challenge remains in the scarcity of high-quality paired video-text datasets, posing a limitation for model training. To address this, several solutions have been explored in the literature, including multi-stage training, image-video mixed training, and mixed-duration video training. These methods allow models to learn more effectively from available data, though training solely on video-text pairs can restrict semantic diversity and sometimes lead to forgetting of image-domain expertise during training [103]. Fine-tuning on higher-quality video subsets has been shown to mitigate these issues, underscoring the need for better video data resources. Another challenge is that most video generation models are trained with a fixed video duration, which necessitates discarding short videos and truncating longer ones, preventing the full utilization of videos with varying frame counts. To address this, mixed-duration training techniques [41] have been introduced. Additionally, some models treat images as still-frame videos, though this often creates a significant domain gap, especially when using bidirectional attention mechanisms.

Looking forward, as vision-language models (VLMs) and multimodal large language models (LLMs) gain influence in video synthesis, they offer new potential to address these limitations and extend the video generation into other applications such as video conversation and video question-answering [104]–[106].

B. Video Editing / Video to Video generation

TABLE I
SUMMARY OF V2V GENERATION MODELS

Model	Tuning-free	Backbone
AnyV2V [107]	✓	Any I2V
Tune-a-Video [108]	x	Stable Diffusion
Video-P2P [109]	x	Stable Diffusion
UniEdit [110]	✓	Any T2V
CoDeF [111]	x	ControlNet
TokenFlow [112]	✓	Stable Diffusion
InsV2V [113]	x	Stable Diffusion
FateZero [114]	✓	Stable Diffusion
Revideo [115]	x	Stable Diffusion
StableVideo [116]	x	Stable Diffusion

Different from video synthesis, in Video-to-Video (V2V) generation, the input is a video in which certain edits are supposed to be applied. Controlling the localization of the edits typically involves additional input guidance such as a segmentation mask, text-based editing prompt, or even another video. Due to the ambiguity of the natural language, text-based video editing is an ill-posed problem as numerous possible edits can satisfy the target text. The masking on the other hand adds additional steps of preparation into the pipeline.

Another significant challenge, similar to the video synthesis domain, is the scarcity of data. *Training* a large-scale video editing model is particularly difficult due to the limited availability of paired data (video-edit instruction) and the extensive computational resources required. To overcome this issue on the data level, using a synthetic paired video dataset has been proposed [113] by taking captions from video and image datasets and using the Prompt-to-Prompt approach and existing T2V models to generate pairs of original and edited videos. *InsV2V* [113] trained on this dataset, is a diffusion-based model that enables video editing using only text editing instruction. Its Long Video Sampling Correction (LVSC) mechanism employs the previous batch's final frames as a reference to guide the generation of subsequent batches, facilitating consistent long video editing. Although the effort of creating such a dataset will be compensated at the inference time as authors avoid per-video-per-model tuning, this approach faces the same limitations as T2V models, reflected in the generated videos.

Hence, to bypass the limitation of the data with other means rather than the data itself, most models follow two common strategies; *zero-shot* adaptation from pre-trained T2I models [112] (training-free) or fine-tuned motion module from pre-trained T2V models [108], [117] (one or *few-shot tuning*). The former approach often suffers from flickering issues because it

lacks a deep temporal understanding, while the latter requires more time and computational overhead to edit videos.

One effective video decomposition and representation approach for video editing is Neural Layered Atlases (NLA). It involves decomposing a video into a series of 2D atlases, each serving as a unified representation of the background or foreground objects across the entire video. Edits made to these 2D atlases are automatically projected back onto the video, ensuring temporal consistency with minimal effort. Based on this technique, *Text2Live* [118] introduces a technique where an edit layer (color + opacity) is composited over the original input, rather than directly generating the edited output. This approach allows Text2Live to achieve high fidelity without relying on user-provided edit masks, unlike most appearance transfer models which are limited to global artistic stylizations or specific image domains. However, atlas representation works best with videos featuring simple motion and is suitable for localized edits driven by straightforward text prompts, such as altering an object's texture or adding complex semi-transparent effects like smoke or fire. Besides, this approach still requires extensive training time.

Employing a pre-trained T2I diffusion model for the task of video editing would alleviate the need for extra training. *StableVideo* [116] combines atlas representation and pre-trained T2I model and employs an inter-frame propagation mechanism. To achieve temporal an aggregation network is designed to generate the edited atlases from the keyframes.

A common approach among training-free models involves applying image editing techniques (e.g., style transfer) on a frame-by-frame basis, followed by a post-processing stage to address temporal inconsistencies in the edited video. For instance, *AnyV2V* [107] simplifies video editing through a *tuning-free* paradigm that operates in two primary steps. First, an off-the-shelf image editing model is used to modify the first frame and next an existing I2V generation model propagates the edits across the entire video by injecting temporal features. This approach supports an extensive array of video editing tasks, including prompt-based editing, reference-based style transfer, subject-driven editing, and identity manipulation. The advantage of this approach is that it can leverage the advancement of a wide range of pre-trained image editing models. Similarly *TokenFlow* [112] offers a *training free* framework that leverages existing models by explicitly propagating diffusion features, based on inter-frame correspondences, readily available in the model. The method exploits the observation that small patches in a natural video extensively repeat across frames and thus consistent editing can be simplified by editing a subset of keyframes and propagating the edit across the video by establishing patch correspondences. This principle holds in diffusion feature space as well, allowing TokenFlow to generate high-quality, text-guided videos that preserve the spatial layout and motion of the original content.

UniEdit [110] presents another **tuning-free** solution for motion and appearance editing using text guidance. It follows an inversion-then-generation pipeline with three branches. The reconstruction branch produces source features for content preservation, and the motion-reference branch yields text-guided motion features for motion injection. The source features and motion features are injected into the main editing branch through spatial and temporal self-attention modules respectively.

Pix2Video [119], offers a training-free video editing technique that begins with a pre-trained structure-guided (e.g., depth) image diffusion model to make text-guided edits on an anchor frame. These changes are then progressively propagated to future frames through self-attention feature injection, adapting the core denoising step of the diffusion model.

FateZero is another zero-shot video editing method without per-prompt training or use-specific mask [114], using source and editing text prompts. At the heart of this model is the Attention Blending Block. All the attention maps in the DDIM *inversion* pipeline are stored and then at the *editing* stage of the DDIM denoising, the Attention Blending Block fuses the editing attention maps with the stored inversion attention maps. More precisely, it replaces the cross-attention maps of unedited words with their attention maps using the source prompt during inversion. While for the edited words, it blends the self-attention maps during the inversion and editing process, with an adaptive spatial mask that represents the areas that the user wants to edit.

Although training-free methods require no heavy training procedure, yet spatiotemporal consistency remains challenging and the models hold the same limitation as their T2I/T2V pre-trained base model. Few-shot tuning is another paradigm that aims at combining zero-shot and training-based models. *Tune-A-Video* [108] motion editing, involves fine-tuning a T2I model to achieve video editing by learning the continuous motion using a tailored spatiotemporal attention mechanism and a one-shot tuning strategy, where only one text-video pair is presented. More precisely, it overfits some diffusion model parameters to a specific video. Then, it uses the overfitting parameters to produce the editing result conditioned on the target prompt.

Video-P2P [109] achieved local editing via video-specific fine-tuning and unconditional embedding optimization. The key innovation of this model lies in the optimization of a shared unconditional embedding for video inversion (inversion of video content into a latent space by text-to-set model), using different guidance for the source and edited prompts, and incorporating their attention maps. However, this model is restricted to only word-swapping prompts due to the reliance on cross-attention.

CoDeF [111] composed of a Canonical Content Field and a Temporal Deformation Field. The former aggregates the static contents of the entire video into a single representation, like a distilled version of the video that captures the essential, unchanging elements across all frames. The latter records the transformations required to convert the canonical image (rendered from the canonical content field) into each frame of the video. CoDeF allows the application of image processing algorithms to the canonical image, which can then be propagated to the entire video using the temporal deformation field. While this model demonstrates high performance in terms of temporal consistency, it struggles with complex scenes involving significant scale changes, sudden emergence of new objects, and multiple fast-moving entities. While the

method aims to improve temporal consistency, there may still be challenges in maintaining perfect consistency across all frames, especially for longer videos or more complex scenes.

Most techniques explained so far, focus primarily on limited local visual content editing while ignoring the motion aspect. *Re-Video* is a video editing [115] model allowing precise control over the visual content and motion within a video, which can be easily extended to multi-area editing without specific training. It encompasses a three-stage training strategy that progressively decouples these two aspects from coarse to fine. A spatiotemporal adaptive fusion module is also introduced to integrate content and motion control across various sampling steps and spatial locations. The regeneration quality though is limited by the base model. Motion control/editing models are discussed more in detail in section III-C2.

In summary, video editing methods generally fall into three main paradigms. *Training-based* approaches are often limited by the scarcity of paired video editing datasets, constraining their applicability in diverse scenarios. *Zero-shot(training-free)* methods entail using pre-trained T2I or T2V models, adapting these for video editing tasks without additional training. *One or few-shot tuning* approaches involve fine-tuning a pre-trained T2I model to generate videos that align closely with desired motions or content. While this tuning demands greater training effort, it offers enhanced flexibility over zero-shot methods. This adaptability is especially valuable in tasks like motion transfer, which will be explored further in Section III-C2. Many current methods are restricted to specific edit types, limiting their flexibility across a broader range of tasks from appearance adjustments, such as style transfer [120] and identity manipulation, to more complex transformations like novel-view synthesis [121].

C. Enhancement techniques

We have analyzed a variety of video generation models and discussed their pros and cons. In this section, we will more especially look into the research works that tried to address existing shortcomings and bring enhancement through their innovations. The improvement can be in terms of temporal or spatial consistency, video motion, and dynamism, optimizing training or inference time, or the duration of the video.

1) ***Spatio-Temporal quality augmentation****:* Given a video sequence $V = \{I_1, I_2..., I_T\} \in \mathbb{R}^{T\times C\times W\times H}$, super-resolution (SR) techniques could be applied along both the temporal and spatial axes. For spatial super-resolution(SSR), the goal is to generate a new sequence $V_{\uparrow SSR} = \{I'_1, I'_2..., I'_T\} \in \mathbb{R}^{T\times C\times W'\times H'}$ while $W' > W$ and $H' > H$ thereby increasing the resolution of each frame. Temporal super-resolution on the other hand, aims to upscale the video along the time axis by a factor of m, producing $V_{\uparrow TSR} = \{I'_1, I'_2..., I'_{T\times m}\}$. It is important to note that TSR results are evaluated also in terms of temporal consistency which considers motion quality and flickering as well.

Traditional video super-resolution methods often rely on fixed degradation models to synthesize training data pairs, which can lead to performance degradation in real-world scenarios. To address this, more advanced data augmentation techniques have been proposed to vary the generation of low-frame-rate and low-resolution video frames from the high-resolution samples. For example, authors in [125], introduce a random downsampling factor m across the temporal axis to generate low-frame-rate sequence $V = \{I_1, I_{1+m}, I_{1+2\times m}..., I_T\}$. Similarly, spatial degradation is simulated by applying a random downscaling factor followed by bilinear interpolation.

Various super-resolution methods exist to enhance spatial, temporal, or spatio-temporal resolution. These methods can be integrated within the video generation process or applied as post-processing [126], [127]. In this study, we focus on embedded super-resolution techniques, possibly offering the advantage of leveraging intermediate features and conditioning inputs from the generation model. For instance, [128] employs optical flow as a conditional input to enhance spatio-temporal quality, by generating temporally coherent optical flow sequences in latent space that are used to warp the input frames. The model consists of a latent flow auto-encoder for spatial content generation and a 3D U-Net-based diffusion model for temporal latent flow generation. These components are trained separately to decouple spatial content generation from temporal dynamics.

Venhancer [125] improves existing T2V generation outputs by enhancing spatial details and synthesizing smoother motion in the temporal domain while mitigating spatial artifacts and flickering. Similar to ControlNet for images, it proposes a conditioning network that injects conditioning features into the base T2V via zero convolutional layers.

VideoElevator [129] is a training-free plug-n-play method that divides the video generation process into two stages. First, in the temporal motion refinement stage, a low-pass frequency filter is applied to enhance the consistency of the video latents which are then processed by a T2V diffusion model to generate natural motion. In the second stage, the denoised latents are deterministically inverted back to noise latents and passed through the spatial quality enhancement stage which uses an inflated T2I model to improve the spatial quality of each frame, resulting in more photorealistic and detailed visuals. Following this design, VideoElevator also supports stylistic customization by integrating personalized T2I models, allowing users to generate videos with specific artistic styles. Similarly, *FreeInit* [130] proposes an iterative refinement of the initial noise during inference to improve temporal consistency. Interestingly they found out that during training, the initial noises corrupted from real videos retain temporal correlation in the low-frequency bands, while i.i.d Gaussian noise used during inference lacks such correlation. To address this gap, they introduce a novel inference-time sampling method that progressively refines the low-frequency components of the initial noise, enhancing temporal consistency and subject appearance without introducing additional learnable parameters.

In summary, we have observed several trends among studies aimed at enhancing spatial and temporal quality within the video generation pipeline. One common approach is to adopt cascaded pipelines where base video generators that produce low-frame-rate and/or low-resolution outputs, are followed

TABLE II
OVERVIEW OF MAINSTREAM VIDEO GENERATION MODELS - LDM: LATENT DIFFUSION MODEL, SO: SAMPLE VIDEOS ONLYS

Model	Modality	Length (s)	Frame rate(fps)	Length (#frames)	Resolution ($W \times H$)	Core technology	Open Code	Open Weight	Open Demo
ImagenVideo [17]	T2V	5.3	24	128	1280×768	Divide & Conquer Diffusion	×	×	SO
Make-a-Video [22]	T2V/I2V	-	-	76	768×768	Divide & Conquer Diffusion	×	×	SO
Phenaki [49]	T2V/I2V	1.4	8	infinite*	1280×720	AutoRegressive Transformer	×	×	SO
AnimateDiff [37]	T2V	2	8	16	$512 \times 512^{\dagger}$	Diffusion Transformer	✓	✓	✓
LaVie [15]	T2V	2	-	16-61 $^{\diamond}$	2048×1280	Divide & Conquer LDM	×	✓	✓
EmuVideo [51]	T2V/I2V	4	16	64	512×512	Brute-force LDM	×	×	SO
Stable Video Diffusion [52]	T2V/I2V	2-5	3-30	14-25	1024×576	LDM	✓	✓	✓
VideoCrafter1 [53]	T2V/I2V	2	8	16	1024×576 1024×640	LDM	×	✓	✓
VideoCrafter2 [54]	T2V	2	8	16	512×320	LDM	×	✓	✓
DynamiCrafter [65]	(I+T)2V	2	8	16	1024×576	LDM	✓	✓	✓
ModelScopeT2V [55]	T2V	2	8	16	256×256	LDM	✓	✓	
CogVideo [18]	T2V	4	8	32	480×480	AutoRegressive Transformer	✓	✓	✓
CogVideox [41]	T2V/I2V	10	16	160	$768 \times 1360^{\triangleleft}$	Diffusion Transformer	✓	✓	✓
Lumiere [14]	T2V/I2V /V2V	5	16	80	1024×1024	Brute-force Diffusion	×	×	SO
Movie Gen [35]	T2V/T2I	16	16	256	1920×1080	Flow Matching Transformer	×	×	SO
I2VGen-XL [64]	(I+T)2V	-	-	32-64	1280×720	LDM	✓	✓	✓
Gen3 [56]	T2V/I2V /V2V	$10^{\triangleright}$	24	-	1280×768	LDM	×	×	✓
Luma Dream Machine [122]	T2V/I2V	$5^{\triangleright}$	24	-	1360×752	Transformer	×	×	✓
KlingAI [59]	T2V/I2V	5-120	30	-	1920×1080	-	×	×	✓
OpenAI Sora [24]	T2V/I2V	60	-	-	1920×1080	LDM	×	×	SO
Open Sora1.2 [11]	T2V/I2V	16	24	-	1280×720	Diffusion Transformer	✓	✓	✓
Pyramid Flow [12]	T2V/I2V	10	24	241	1280×768	Flow Matching Transformer	✓	✓	✓
Mochi1 [123]	T2V	5.4	30	-	640×480 1280×720	Diffusion Transformer	×	✓	✓
Allegro [124]	T2V	6	15	88	1280×720	VAE Transformer	×	✓	✓

* Theoretically infinite (autoregressive)
† Follows T2I base model
$^{\diamond}$ By interpolation
$^{\triangleleft}$ Open-sourced version generates 6s, 720×480 @ 8fps
$^{\triangleright}$ Extendable by 5s intervals

by temporal and spatial upsampling blocks [15], [17], [38], [64]. Another approach incorporates temporally coherent conditional signals, such as optical flow [128], [131] during generation to enhance motion consistency and temporal alignment between frames. Additionally, plug-in modules are becoming popular for leveraging pre-trained video generation models and avoiding resource-intensive retraining. These modules can be classified into two main groups: those requiring limited training or fine-tuning, such as ControlNet-inspired methods [125], and training-free methods [129], [130] which provide flexible, plug-and-play enhancements without additional training or fine-tuning required.

*2) **Motion control**:* In recent years, customization of video generation via various conditioning inputs has been extensively explored to address the need for finer control over object appearance and motion. While text and image signals provide strong guidance for appearance, they are often insufficient to express temporal aspects of video, such as camera movements or complex object trajectories. To overcome this, researchers have experimented with a range of conditioning inputs, from sparse signals like sketches and trajectories to dense inputs such as masks, human poses [132], and depth maps. For instance, *SparseCtrl* [74] is a plug-n-play controller that injects temporally sparse control signals (e.g., sketch, depth, RGB image) into the diffusion process, using an additional encoder on top of the base T2V model, enabling the base model to be used for various applications, such as sketch-to-video, depth-guided generation, and image animation. *VideoComposer* [133] extends this by enabling the composition of various modalities, offering even greater control over generated videos.

Camera motion control: A simple way to model camera motion in 3D space is through a camera movement vector, denoted by (c_x, c_y, c_z) which corresponds to x-pan, y-pan, and zoom ratio, respectively. This has been simplified to a 2D displacement of the camera in some papers. A more sophisticated approach involves using the camera pose, which in computer vision refers to the position and orientation of the camera in three-dimensional space relative to a world coordinate system. Mathematically it is represented by intrinsic matrix $K \in \mathbb{R}^{3\times3}$ and extrinsic matrix $E = [R; t]$, where $R \in \mathbb{R}^{3\times3}$ describes the camera's rotation, and $t \in \mathbb{R}^{3\times1}$ represents the translation vector. Most video generation models lack precise control over camera viewpoints, making it challenging to adjust or simulate camera motion effectively. Various methods have been explored to introduce camera motion control by incorporating different control or conditioning signals at multiple levels, including basic motions (e.g., zoom, pan), hybrid motions, or complex trajectories. Typically, these signals are fed into an additional encoder, which is then injected into the video generation model. However, controlling camera motion presents

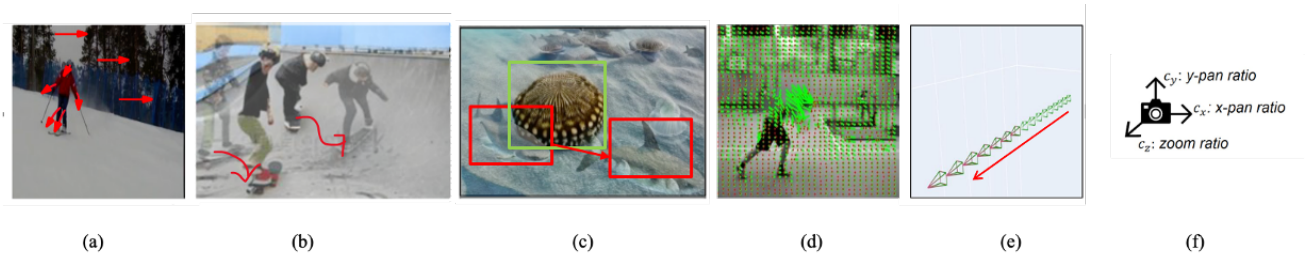

Fig. 6. Examples of input signal for object motion guidance: (a) Strokes (b) Point trajectory (c) Bounding box trajectory (d) Optical flow. Examples of input signals for camera motion guidance: (e) Camera pose[R,t] (f) 2D/3D displacement vector

several challenges, such as the scarcity of annotated data, particularly data that contains precise descriptions of camera movements. Moreover, due to the complexity of training models, it is desirable for camera motion to be seamlessly integrated into existing video generation frameworks without compromising frame quality or temporal consistency.

One strategy to address the data limitation is through data augmentation [134], which simulates camera movements in videos. However, relying solely on this approach limits the scope of control to simple motions such as zooming or trucking .

Another promising approach involves fine-tuning plugin modules on specific types of camera motion. For example, *AnimateDiff* [37] introduces an efficient fine-tuning method known as MotionLoRA, which allows for personalized motion control. MotionLoRA layers can be fine-tuned using as few as 20–50 reference videos over 2,000 training iterations. These layers' low-rank structure enables composability, allowing for combinations of individually trained MotionLoRA models to produce complex, composite motion effects. Similarly, *MotionDirector* [135] controls motion by extracting it from one or more template videos, using temporal and spatial LoRA modules.

Another common method is to embed camera parameters into diffusion models via learnable encoders, followed by extensive fine-tuning on large-scale datasets containing detailed camera trajectories. For instance, *CameraCtrl* [136], utilizes a Plücker embedding to associate each pixel in a frame with both the camera's center and the direction vector pointing from the camera to that pixel. This embedding facilitates comprehensive camera pose descriptions and easier learning, as its components tend to have uniform value ranges Although the camera features can be easily integrated into pre-trained T2V models via temporal attention layers, the approach is still limited by the need for a diverse dataset with complex camera trajectories. A training-free alternative is *CamTrol* [137], a plug-and-play camera control module that follows a two-stage process. First, explicit camera movements are modeled using 3D point cloud representations to render a series of images based on a predefined camera trajectory. Then, a layout prior derived from noisy latents—obtained via the diffusion inversion process on the rendered images—guides the generation of videos with camera movements. While this method avoids the need for training, it introduces a trade-off between the fidelity and diversity of generated videos, due to the limitations in separating appearance and motion in the noise latent space. In general, Noise distribution plays a critical role in the dynamism of generated videos. In diffusion-based image-to-video (I2V) models, conditional image leakage is a notable issue, where models over-rely on the input image, leading to videos that lack dynamic motion. To address this, authors in [138] propose two strategies: 1) a training-free inference method that initiates the generation process at an earlier time step, to avoid the unreliable late-time steps of I2V diffusion models, 2) an initial noise distribution with optimal analytic expressions. Additionally, a time-dependent noise distribution can be introduced during training to interfere with the conditional image, ensuring that high noise levels at later time steps sufficiently disrupt the input image and encourage the model to generate more dynamic video content.

Object motion control: *MCDiff* [139] controls the motion through sparse flow (strokes) inputs. Its flow completion model first predicts dense flows representing that represent per-pixel momentary motion. The model synthesizes future frames autoregressively, based on both the previous frame and the predicted dense flow, in a conditional diffusion process. Nevertheless, the model can only control motion from humans as it relies on human body key points extraction for each person to construct data. *MotionI2V* [131] focuses on first predicting the plausible motions in the form of pixel-wise trajectories by tuning a pre-trained video diffusion model for optical flow prediction using an input image and textural description. The predicted 2D displacement maps, which contain the optical flow between the reference frame and future frames, are then used to warp the features of the reference frame. These features are injected into the synthesized frames through cross-attention mechanisms, thereby enhancing the temporal receptive field. Additionally, the authors introduce sparse trajectory guidance, which is refined using their proposed trajectory ControlNet for more accurate motion prediction. *DragAnything* [140] challenges the assumption made in previous works [141], [142], that a single point on the target is sufficient to represent the target for motion prediction. A novel Entity Representation method is proposed to achieve precise motion control for any entity in a video, using the latent features of a diffusion model. First, the latent noise of the input image is obtained through diffusion inversion. Then, a denoising U-Net is employed to extract the corresponding latent diffusion features. Having these features and the entity mask, the entity embeddings are retrieved by indexing the appropriate coordinates. Finally,

these entity embeddings get associated with the corresponding trajectory points, and the encoders are used to encode them into the latent feature space. It is important to note that trajectory-based motion control is inherently limited to 2D, making it difficult to handle more complex 3D motion, such as precise body rotations or movements involving turning around.

Camera and object motion control: To achieve finer control over generated content, some methods focus on the simultaneous control of both object and camera motion. One such approach is *Direct-a-Video* [134] which decouples object motion from camera movement. It introduces a camera embedder to control the camera pose during video generation and a training-free spatial cross-attention modulation for objects' motion control. Although this model offers control over both the camera and multiple objects in a scene via bounding boxes, it is limited to conditioning only three camera parameters. As a result, the complexity of the camera trajectory is constrained to basic movements such as panning and zooming. *MotionCtrl* [142] Incorporates a camera and object motion controllers that accept a sequence of camera poses and object trajectories as input, generating features that are then injected into the temporal and convolutional modules of the video generator. However, the need to fine-tune part of the video diffusion model can reduce its generalization capabilities.

DragNUWA [141] stands out by handling complex curved trajectories and managing multiple object movements alongside camera motion simultaneously. It achieves this by warping images through a combination of densified sparse strokes and pixel fusion. To address the challenge of limited trajectory ground truth data, DragNUWA introduces a Trajectory Sampler that directly samples trajectories from video optical flow, enabling the model to learn a wide range of possible trajectories in an open-domain setting. To address the challenge of limited trajectory ground truth data, DragNUWA introduces a Trajectory Sampler that directly samples trajectories from video optical flow, enabling the model to learn a wide range of possible trajectories in an open-domain setting. Likewise, *VideoComposer* [133] leverages MPEG-4 to extract motion vector information from videos, using these vectors as conditions during training. However, due to the lack of high-level semantic information within motion vectors, this method only allows for reproducing simple object movements.

Boximator [143] proposes a plug-and-play control module that simplifies the challenge of long-range spatial-temporal information propagation by factorizing it into two more manageable tasks. First, the model generates a bounding box for each object with a dedicated color. Second, it aligns these boxes with Boximator constraints in each frame. Initially, the model is trained to generate videos with visible colored bounding boxes around objects; in a subsequent phase, it is further trained to stop producing these visible boxes, while preserving the motion. Additionally, Boximator introduces the concept of *soft boxes* for flexible object shaping and motion paths, alongside *hard boxes* for precise object positioning and shape definition.

Video-to-Video motion transfer (one-shot customization): Another way to bring the motion to video generation is by motion transfer from another video or sequence as a prior [144]–[146]. For instance, *MotionClone* [144] is a training-free framework that enables the cloning of motion from a reference video to control T2V generation. This method allows for motion transfer across different object categories while preserving essential motion characteristics. However, the authors highlight limitations in out-of-distribution scenarios, where the combination of target objects and input video motions may lead to visual artifacts. Similarly, authors in [145] utilize a pre-trained, fixed T2V diffusion model combined with a novel space-time feature loss, derived directly from the model, ensuring that the generated video adheres to the overall motion of the input while complying with the target object in terms of shape and fine-grained motion trait.

Control-A-Video [117] infuses motion priors from a reference video through residual-based and optical flow-based noise initialization, promoting coherence among frame latents, thereby reducing flickering across frames. Additionally, this model incorporates a Spatio-Temporal Reward Feedback Learning algorithm, which optimizes the video diffusion model using multiple reward functions to enhance video quality and motion consistency. Another noteworthy approach is Customize-A-Video [146], which adapts motion from a single reference video to new subjects and scenes by employing Low-Rank Adaptation (LoRA) on temporal attention layers. The method first trains an appearance absorber module on unordered reference frames to capture spatial information from each frame. Afterward, with the help of the trained appearance absorber, a Temporal LoRA module is trained to focus specifically on the motion from the reference video. During inference, the appearance absorber is discarded, and only the trained Temporal LoRA module is used for motion transfer.

VMC [147] offers a different approach by distilling motion trajectories from the residual between noisy latent frames. It fine-tunes only the temporal attention layers of the keyframe generation model based on these motion trajectories. Once training is complete, the customized key-frame generator is leveraged for target motion-driven video generation with new appearances. To enhance this process, the model uses an appearance-invariant prompt to filter out background information that might interfere with motion extraction. However, VMC struggles to generalize when the appearance of the target object differs significantly from the reference object. A common challenge faced by these one-shot methods is their relatively high inference time effort, as each video requires specific model tuning. While these techniques excel at customizing motion for individual videos, they can be computationally demanding, especially when applied to diverse and complex video generation tasks.

Video Personalization: Subject customization in video generation is typically achieved through either a few-shot fine-tuning of the model (using images) or textual inversion (via a learnable text embedding), without the need for model fine-tuning. *MotionBooth* [148] employs a few-shot approach to fine-tune a T2V model, allowing it to capture the object's shape and attributes. Additionally, it introduces a training-free technique that manipulates cross-attention maps to control subject motion, along with a novel latent shift module for

TABLE III
SUMMARY OF MODELS AND TECHNIQUES AIMING AT MOTION CONTROL

Model	Camera control	Object control	Training requirement	Year
Tune-a-Video [108]	using a reference motion video	using a reference motion video	one-shot tuning	2023
Video-P2P [109]	using a reference motion video	using a reference motion video	one-shot tuning	2024
AnimateDiff [37]	composition of basic movements	×	required for MotionLoRA module	2024
CameraCtrl [136]	camera pose	×	required for control module	2024
CamTrol [137]	camera pose	×	zero-shot	2024
Direct-a-Video [134]	composition of basic movements	bounding boxes	required	2024
MotionCtrl [142]	camera pose	trajectory	required for motion modules	2024
MotionDirector [135]	using a reference motion video	using a reference motion video	required for LoRA module	2023
DragNUWA [141]	implicitly by custom trajectories	trajectory	required	2023
Control-A-Video	using a reference motion sequence	using a reference motion sequence	required	2023
DragAnything [140]	basic motions	trajectory (bg and fg control)	required	2024
Boximator [143]	implicitly by custom trajectories	bounding boxes/trajectory	required for control module	2024
VideoComposer [133]	reference motion video/strokes	reference motion video/strokes	required	2024
MotionI2V [131]	×	text/sparse trajectory	required	2024
MotionClone [144]	using a reference motion video	using a reference motion video	zero-shot	2024
Space-Time Diffusion [145]	using a reference motion video	using a reference motion video	zero-shot	2024
Customize-A-Video [146]	using a reference motion video	using a reference motion video	one-shot tuning	2024
VMC [147]	using a reference motion video	using a reference motion video	one-shot tuning	2024
MotionBooth [148]	camera displacement	bounding boxes	one-shot tuning	2024
DreamVideo [149]	using a reference motion video	using a reference motion video	few-shot tuning for adapter module	2023

camera movement control.

Inspired by Dreambooth [132] -a personalized image generation method that binds a word to a subject through complete fine-tuning of an image diffusion model- *DreamVideo* [149] extends this concept to video generation. DreamVideo uses a few-shot approach, creating personalized videos from a small set of static images of the desired subject and a few videos illustrating the target motion. The task is decoupled into two distinct phases of subject and motion learning. In the subject learning phase, the model captures the subject's fine details from the provided images by fine-tuning an identity adapter module. In the motion learning phase, a motion adapter is fine-tuned to model the target motion pattern effectively. These lightweight adapters, together with a randomly selected image of the target (used as appearance guidance), are used on top of a frozen video diffusion model to generate customized videos.

In contrast to DreamVideo, which is limited by predefined motion types and lacks flexibility in handling text-driven input, *MotionBooth* offers greater control over both subject and camera motions without relying on predefined motion prototypes. Similarly, *VideoBooth* [150] generates videos based on subjects specified in image prompts, using an image encoder that integrates the image prompts into text embeddings, mapping them to multi-scale latent representations. This controls the generation process via cross-frame attention layers within T2V models. However, the generalization capacity of VideoBooth to diverse subjects, such as human figures, remains limited.

To summarize, most video generation models still lack precise control over motion and temporal consistency, particularly in complex scenarios involving both camera and object motion. Furthermore, traditional 1D temporal attention mechanisms often fail to capture long-range temporal dependencies due to their narrow receptive field, resulting in inconsistency, especially in the presence of large motion. Researchers have explored various paradigms to address these challenges, listed in Table III:

- Customized motion by data augmentation and plug-and-play modules: Simple camera motions can be learned through data augmentation [134] or plug-and-play modules like MotionLoRA using few-shot learning to mimic specific movements [37].
- Motion Transfer: Methods that transfer motion from reference videos offer another solution, although real-world applications are limited by the difficulty of obtaining dense motion guidance.
- Motion Feature Injection: By injecting motion priors—such as dense optical flow or sparse point trajectories—into video generation models, finer control over motion can be achieved. Sparse point trajectories are commonly used for both implicit camera motion control and explicit object manipulation, typically in two paradigms: Trajectory Map (point) and bounding box representation. The challenge with using point trajectories is that a single point sometimes fails to adequately represent an entire entity. Additionally, pixels closer to the drag point tend to receive a greater influence which sometimes leads to deformation in appearance. The box representation, on the other hand, is limited to instance-level objects and cannot account for backgrounds. An alternative approach involves extracting entity-level latent features and incorporating them into the video generation process. Another promising strategy is embedding camera pose or trajectory directly into the diffusion process [136], [142].
- Video Personalization: Few-shot fine-tuning or textual inversion techniques (through a learnable text embedding without model fine-tuning) allow for the customization of a given subject's motion and appearance.
- Latent Space Optimization: Manipulating the initialization of latent distributions, as seen in training-free approaches [137], [144], can enhance video dynamism without requiring extensive fine-tuning. This technique is also applicable for fine-tuning the video generation model.

*3) **Long duration**:* Video generation output is characterized by interleaved factors such as frame count, framerate, and duration (since $framecount = duration(s) \times framerate(fps)$). Conventionally, videos are considered *long* if they contain more than 100 frames or equivalently, or if their duration exceeds 10 seconds at a framerate of 10fps [9]. Generating long videos presents several challenges. One key difficulty is preserving temporal consistency and continuity in the synthesized motion while maintaining realism, consistency in the appearance of entities, and limiting computational resource usage. Data scarcity for long video datasets adds further complications. Additionally, training models on long videos is extremely expensive, and without sufficient training, the quality degrades during inference. This is especially problematic for attention blocks, which are typically trained to focus on a limited number of neighboring frames. Some approaches have been inspired by the large context capabilities of large language models (LLMs). For instance, ExVideo [36] proposes to fine-tune the parameters within the temporal layers of a pre-trained T2V model, as well as incorporating learnable positional embeddings to handle larger contexts. This strategy enables the temporal layers to process longer temporal context and generate up to 128 frames. Despite these enhancements, the approach remains constrained by the inherent limitations of its base T2V model.

To address these challenges, several techniques have been proposed, primarily following two paradigms: *autoregressive* generation and *divide-and-conquer* generation. In the divide-and-conquer approach, the model first generates keyframes that define the main narrative and then fills the gaps between these keyframes. This method can take advantage of parallelization for the infilling process, which accelerates the overall video generation. However, maintaining coherence between chunks of the video across different keyframes remains a challenge.

On the other hand, the autoregressive paradigm sequentially generates small chunks of video, each conditioned on the previous frame(s) (and/or its clip embedding) of the previous chunk(s). Conditioning solely on the last frame can lead to video stagnation and longer conditioning could cause the error accumulation effect. Ideally, long-term memory (for the appearance of entities) and short-term memory (for temporal dynamics) are needed to create diverse motions and narratives. This is often achieved through conditioning mechanisms or mask modeling. Masked visual modeling is a technique to selectively obscure parts of video frames to enhance the model's learning process. The model then learns to predict these masked parts based on the visible context and the sequence's temporal dynamic. For example, authors in [38] propose using probabilistic masks based on the Bernoulli distribution or predetermined patterns to selectively obscure parts of input frames during training.

Figure 7 illustrates several examples of long video generation conditioning schemes. One simple method (Figure 7-a)involves using temporal sliding windows so that the temporal attention module can consistently process a fixed number of frames. For instance, Gen-L-Video [154] generates long videos by merging overlapping chunks using a sliding-window method during denoising. It treats long videos of arbitrary lengths and multiple semantic segments as collections of short videos with temporal overlap, resulting in smooth temporal transitions, yet it struggles with maintaining long-range visual consistency. In contrast, authors in [155] introduce a more flexible generative model that can sample any arbitrary subset of video frames, conditioned on any other subset during inference time. This permits the choice to extend videos either autoregressively or via a hierarchical generation(Figure 7-d).

Divide-and-conquer approaches: NUWA-XL [16] employs a "Diffusion over Diffusion" architecture to generate long videos. A global diffusion model first creates the storyline by generating L keyframes based on L prompts. Then, local diffusion models fill the gaps between these keyframes. One significant advantage of this setup is that it enables training on long videos, eliminating the training-inference gap. Typically, models are trained on fixed-length clips as small as 16 frames and forced to extend generation to larger scales, leading to a domain gap. However, this approach requires extensive pretraining on large long-video datasets and demands a performant global diffusion model to produce initial keyframes. Moreover, its generalization capability for open-domain video generation remains untested due to the limited dataset used.

Autoregressive approaches: StreamingT2V [151] incorporates a long-short-term conditioning mechanism built on a frozen T2V model. It consists of three main components: a conditional attention module as a short-term memory, an appearance preservation module as a long-term memory, and a randomized blending approach for seamless blending of overlapping video chunks. Each frame is synthesized based on features extracted from the previous chunk using an attention mechanism to ensure smooth transitions. NUWA-infinity [19] takes a different approach by splitting long visuals into non-overlapping patches, using an ordered patch chain as a complete training instance. A rendering model then autoregressively predicts each patch based on its context. [157] proposes Frame-Level Noise Reversion to reuse the initial noise from previously generated clips. This helps preserve temporal coherence as the reverse sequence is still temporally consistent, yet promotes visual diversification, avoiding frame-level jittering and disjointed transitions. FIFO-Diffusion [156] employs training-free iterative diagonal denoising to generate infinitely long videos using a pre-trained video generation model. This diagonal denoising process applies increasing noise levels to consecutive frames, unlike traditional methods that use uniform noise levels. Interestingly, FIFO-Diffusion maintains a constant memory footprint regardless of video length, making it ideal for parallel inference on multiple GPUs. FreeNoise [152] is a training-free approach that reschedules noise sequences for long-range correlations instead of initializing noises for all frames at once. It is challenging for the temporal modules to achieve global coherence when independently sampled noises are combined for longer video generation. Therefore, temporal attention is performed over these noise sequences using a window-based fusion method. Despite their novel motion injection method to support multi-prompt conditional generation, the model tends to produce near-static global motion in long videos.

TABLE IV
OVERVIEW OF LONG VIDEO GENERATION MODELS. #FRAMES IS PROVIDED BASED ON THE LONGEST PLAUSIBLE EXAMPLES PROVIDED BY AUTHORS

Model	Mode	#Frames	Resolution	Tech
StreamingT2V [151]	T2V	1200+	720×720	AutoRegressive
NUWA-XL [16]	T2V	3376+	256×256	Divide-and-conquer
Phenaki [49]	T2V/I2V	2+ minutes	-	Autoregressive
FreeNoise [152]	T2V	512	follows T2V base model	Autoregressive
Vlogger [153]	T2V/I2V	5+ minutes	320×512	Divide-and-conquer
Gen-L-Video [154]	T2V	hundreds	follows T2V base	Hybrid
FDM [155]	V2V	15000	128×128	Hybrid
FIFO-Diffusion [156]	T2V	1000	follows T2V base model	AutoRegressive

Multi-Scene / prompt generation: One limitation of models like FIFO-Diffusion [156] is that although they produce long and relatively consistent videos, the storyline remains static. A single text condition is often inadequate and ambiguous to fully describe evolving content. To address this, some works have explored multi-scene long video generation [153], [158], [159]. Ideally, a model should maintain the identity of entities throughout the entire video while varying their actions, background, and story. For example, VideoDirectorGPT [159] offers explicit control over spatial layouts and maintains temporal consistency of entities across multiple scenes. This framework integrates knowledge from large language models (LLMs) for video content planning and grounded video generation, expanding initial text prompts into a detailed *video plan* including scene descriptions, entities, background details, and consistent groupings. Similarly, Vlogger [153] converts user stories into scripts through rounds of interaction with an LLM. Based on this script, the model generates actor reference images using a T2I model and assigns actors to scenes. MEVG [160] employs a last-frame-aware diffusion process to preserve visual coherence between consecutive videos. Each video consists of different events based on the text generated by the prompt generator LLM and a pre-trained T2V. It simultaneously adjusts noise in the latent to enhance the motion dynamic in a generated video.

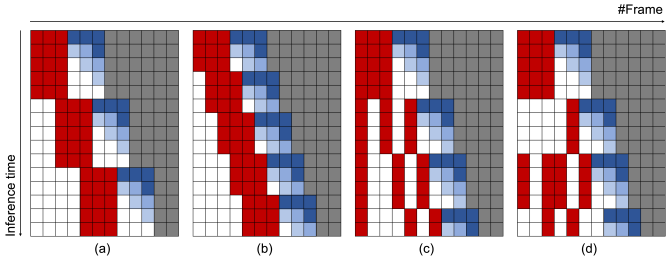

Fig. 7. Examples of long video generation condition schemes. a: Chunk Auto-regressive, b: FIFO Auto-regressive [156], c: Long range Auto-regressive, d: Flexible Auto-regressive [155] Red represents conditioning frames(already generated), and the blue frames are gradually getting denoised(white) while gray frames wait for the next steps to get denoised.

*4) **Training/Inference Efficiency**:* The efficiency of the training and inference pipeline is a key challenge, particularly for video generation models, which are often resource-intensive. This issue is even more pronounced for diffusion-based models which rely on iterative inference pipelines. Efforts to optimize efficiency span several aspects of the video generation process:

- **Data:** The quality of training data plays a crucial role in the final output excellence. Filtering the dataset to include high-quality data has been shown to enhance model performance. For instance, as recommended in [52], [161], excluding motionless clips or those with excessive on-screen text improves the model's performance. Motionless clips could be identified using optical flow relying on the significant correlation between a video's optical flow score and motion intensity, with low scores indicating static frames and high scores indicating intense emotion. Likewise, optical character recognition techniques could be applied to identify the clips containing large amounts of written text (e.g. watermarks).
  However, high-quality training videos are scarce and when training with fixed-duration videos, short clips may need to be discarded, and longer videos truncated, leading to under-utilization of data with varying frame counts. To address this, the authors of [41] propose a mixed-duration training strategy called Frame-Pack, which allows the use of variable-length videos during training. This approach eliminates the need to trim or discard videos, enabling better utilization of data and improving the model's generalization capabilities. Additionally, most existing datasets lack detailed video captions. Captions are often limited to high-level descriptions such as "A dog running" without including temporal or detailed compositional information. Enriching video captions using Vision-Language Models (VLMs) can generate more granular descriptions, such as detailed body movements [162]. This technique not only enhances caption diversity but can also be used for data augmentation.
- **Training strategy**: Progressive training is a method where the spatial and/or temporal resolution of the videos increases as training progresses [41], [51]. For example in EmuVideo [51], most of the training occurs on 1-second long 256×256 videos at 8 fps, reducing per-iteration training time by 3.5x. The model is then progressively trained on 2- and 4-second long, 512×512 resolution videos for fewer iterations. Progressive training allows the models to learn coarse-grained details early and fine-grained details later through high-resolution training, all while reducing overall training time and effectively utilizing videos of various resolutions. Another strategy involves pre-training models on large-scale text-image

datasets before fine-tuning on video datasets. This allows for leveraging large high-quality image datasets publicly available, to improve video generation with minimal reliance on video data. Additionally, unsupervised training frameworks can further reduce dependency on paired text-video data. For example, TF-T2V [163] is a plug-and-play paradigm that employs separated content and motion branches, with the former learning text/image conditional appearance generation from image-text pairs and the latter synthesizing motion dynamics from text-free videos.

- **Quantization**: Mainstream quantization methods can be categorized into two groups: quantization-aware training(QAT) and post-training quantization(PTQ). QAT incorporates quantization simulation during the training phase to preserve performance under reduced precision. However, this approach necessitates substantial computational resources, extended training time, and access to the original dataset. In contrast, PTQ does not require fine-tuning and can be implemented with minimal computational effort, using only a limited set of unlabeled data for calibration. PQT is an effective technique for reducing memory usage, computational complexity, and latency in video generation models. Compressing high bit-width floating-point data into lower bit-width integers helps accommodate lower-memory GPUs while maintaining minimal video quality degradation [164]–[166].
- **Model design**:
  - Diffusion process optimization: The v-parameterization technique improves stability during inference with fewer sampling steps and it is useful for avoiding color shifting artifacts, particularly for high-resolution diffusion models [17]. In v-parametrization, instead of predicting the noise ϵ_t or the clean data x_0 directly, the model predicts an intermediate latent variable v which is designed as a blend of both the clean image x_0 and the Gaussian noise ϵ_t at timestep t, striking a balance between the two: $v_t = \alpha_t \epsilon_t + \sigma_t x_0$
    Progressive distillation [167] is another technique that accelerates diffusion models by iteratively reducing the number of sampling steps. This process distills a high-step "teacher" model into a "student" model that can generate samples with fewer steps. To enhance the distillation output, T2V-Turbo [168] further integrates reward feedback from multiple models during distillation, optimizing human preference and the temporal coherence of the generated video.
  - Transformer optimization: CogVideoX [41] utilizes a video-adapted version of Rotary Position Encoding (RoPE-3D), a relative positional encoding method proven to capture inter-token relationships effectively in large language models particularly excelling in modeling long sequence. 3D-RoPE accelerates model convergence compared to traditional sinusoidal encodings.
  - Lightweight plug-and-play modules: To avoid the computational burden of extensive fine-tuning, plug-and-play modules offer an alternative. For instance, MotionLoRA from animatediff [37] enables efficient tuning of text-to-video models without modifying the base model architecture. LoRA is a technique for accelerated fine-tuning of large models. Instead of updating the entire weight matrix of a model $\mathcal{W} \in \mathbb{R}^{m \times n}$ during fine-tuning, LoRA decomposes the weight updates into smaller, lower-rank matrices and optimizes only these newly introduced matrices:

$$\mathcal{W}' = \mathcal{W} + \Delta\mathcal{W} = \mathcal{W} + AB^T \tag{18}$$

    where $A \in \mathbb{R}^{m \times r}$ and $B \in \mathbb{R}^{n \times r}$ are a pair of rank-decomposition matrices, r is the rank of LoRA layers.

IV. Evaluation metrics

Video quality evaluation methods can be classified into two principal categories: quantitative (objective) and qualitative (subjective) approaches. Qualitative assessments typically involve human evaluation, where groups of evaluators score videos based on various criteria such as photorealism, text alignment, temporal coherence, and aesthetic appeal. Despite the labor-intensive nature of human evaluation, it remains indispensable due to the human capacity to recognize temporal coherence, interpret ambiguity, and understand real-world physics—capabilities that quantitative metrics often fail to fully replicate due to their inherent limitations in frame-level comprehension [35].

Quantitative evaluation methods can be further categorized into three primary types: text-video alignment, spatial quality, and temporal quality assessment. Additionally, these metrics can be classified based on their reliance on ground truth (paired or set), distinguishing between standalone(unary) and comparative metrics.

A. Text-to-video alignment

This category of metrics assesses the degree to which generated videos correspond with their associated textual descriptions/prompt. CLIP-based methods, including CLIP [170], BLIP [171], and viCLIP [103] score are prevalent in the literature. These models are trained to maximize the similarity between pairs in extensive text-image/video datasets. However, these approaches often exhibit inconsistencies with human visual perception, leading to increased interest in Learning from Human Feedback (LHF). This has encouraged the creation of large human-rated datasets for alignment purposes. Consequently, new alignment models such as ImageReward [172], PickScore [173], and VideoScore [174] have emerged, aiming to facilitate automatic quantified video quality assessments that maintain a high correlation with human preference.

B. Spatial quality assessment

Common metrics for measuring image quality include *Inception Score (IS)* and *Fréchet Inception Distance (FID)*. However, these metrics often demonstrate a weak correlation

TABLE V
COMPARISON OF HARDWARE REQUIREMENTS FOR VARIOUS STATE-OF-THE-ART VIDEO GENERATION MODELS. THE TRAINING HARDWARE SPECIFICATIONS COLUMN DETAILS THE COMPUTATIONAL POWER USED BY AUTHORS TO TRAIN THE PUBLISHED MODEL WEIGHTS, OR THE MINIMUM HARDWARE NEEDED TO INDEPENDENTLY TRAIN OR FINE-TUNE EACH MODEL.

Model Name	Training Hardware Specification	Inference Hardware Specification	#Parameters
T2V-Turbo-v2 [168]	8 × A100 GPU for 10K	—	—
VEnhancer [125]	16 × A100 GPU 80GB VRAM for 4 days	Minimum A100 80G VRAM	—
Pyramid Flow [12]	A100 GPU for 20.7K hours (minimum 8 × A100 GPU needed)	Less than 8GB VRAM† ◇	2B
CogVideoX-5B [41]	1 × 4090 GPU 24 GB VRAM	18-26GB VRAM (Desktop GPUs like RTX 3060)	5B
Allegro [124]	256 × H100 for total of 252k iterations in different phases	1 × GPU 9.3GB VRAM † ▷	VAE: 175M DiT: 2.8B
CogVideoX-2B [41]	47-62 GB VRAM required	4-18GB VRAM	2B
VideoCrafter-2.0 [54]	32 × A100 GPU for 270K iterations (Training) + 8 × A100 GPU for 30K iterations (Finetuning)	—	—
OpenSora V1.2 [11]	H100 GPU for 35K hours	More than one GPU 80G VRAM	1.1B
Mochi1 [169]	—	Minimum 4 × H100 GPU	VAE: 362M DiT: 10B
Movie Gen [35]	Up to 6144 × H100 GPU	—	30B*
Stable Video Diffusion [52]	8 × A100 GPU 80GB VRAM for 12K iterations(16 hours)	—	1.5B

† With CPU offloading
◇ To generate a 5s, 768p, 24fps video, takes 5.5 minutes on 1 × A100 GPU, or 2.5 minutes on 4 × A100 GPU
▷ To generate a 6s, 720p, 15fps video, takes 20 minutes on 1 × H100 GPU, or 3 minutes on 8 × H100
* Transformer only (other modules such as text embedded or TAE not included)

with human visual perception. Other metrics frequently employed for image-level video assessment include *Peak Signal-to-Noise Ratio (PSNR)* and *Structural Similarity Index (SSIM*). PSNR quantifies the peak signal-to-noise ratio relative to Mean Squared Error, while SSIM evaluates differences in brightness, contrast, and structure between reference and generated videos. Although these metrics were initially developed for image tasks such as super-resolution, they have been adapted for video evaluation without necessitating pre-trained models. Newer metrics, such as *UNIQUE* [175] and *MUSIQ* [176] have shown potential in effectively capturing perceptual quality in natural images.

C. Temporal quality assessment

Temporal quality is primarily evaluated using metrics such as *Fréchet Video Distance (FVD)*, which measures feature disparities between generated and real videos through the application of Inflated-3D ConvNets (I3D). Additionally, *Kernel Video Distance (KVD*) evaluates the quality of generated videos by comparing the distributions of real and generated videos within a feature space, utilizing kernel functions (e.g., Gaussian kernel) to measure similarity. While FID, FVD, and KVD compare the distribution of features of generated frames against real images/videos, they may overlook distortion-level and semantic-level quality characteristics. The *Fréchet Video Motion Distance (FVMD)* specifically focuses on evaluating the motion quality. Unlike FVD, which assesses both spatial and temporal quality more broadly, FVMD emphasizes the dynamics of motion, capturing the coherence, fluidity, and realism of object movement across frames. *Deep Objective and Visual Evaluation for Robustness (DOVER)* [178] assesses video quality from two perspectives: the aesthetic aspect, which considers content, composition, and other non-technical factors, and the technical aspect, which focuses on the perception of distortions and technical characteristics such as blur and artifacts.

D. Leaderboard benchmarks

To facilitate fair comparisons across video generation models on a common basis, several benchmarks have emerged. These benchmarks often include a variety of text prompts covering distinct concepts (e.g., different subjects, landscapes, and motion levels) [35], [179]–[181]. Commonly used benchmarks include VBench [182] and EvalCrafter [183] each providing extensive datasets of text prompts, as well as a combination of various evaluation metrics to evaluate *T2V* models with a final scalar score.

VBench proposes a comprehensive set of fine-grained video evaluation metrics to assess temporal and spatial video quality, as well as video-text consistency in terms of semantics and style, using a list of 800 prompts. It decomposes video generation quality into 16 dimensions such as subject identity inconsistency, motion smoothness, and temporal flickering, and proposes the evaluation metrics with fine-grained levels. Figure and Table VII illustrate the performance of several state-of-the-art models on VBench. Notable performance differences emerge in metrics like multiple object handling and dynamic degree, while the models demonstrate similar performance levels on metrics such as background consistency. EvalCrafter benchmark consists of 700 prompts for T2V generation. The generated videos are assessed in terms of visual qualities, content qualities, motion qualities, and text-video alignment using 17 selected objective metrics. A human alignment method is used to find the best coefficients to combine those metrics instead of simply averaging. I2V-Bench [180] is a comprehensive evaluation benchmark for Image-to-Video (I2V) generation models, features 2,950 curated YouTube videos based on strict resolution and aesthetic standards and organized across categories like scenery, sports, animals, and portraits.

Although existing benchmarks primarily address temporal consistency and continuity, they frequently neglect content dynamics, essential for evaluating visual vividness and align-

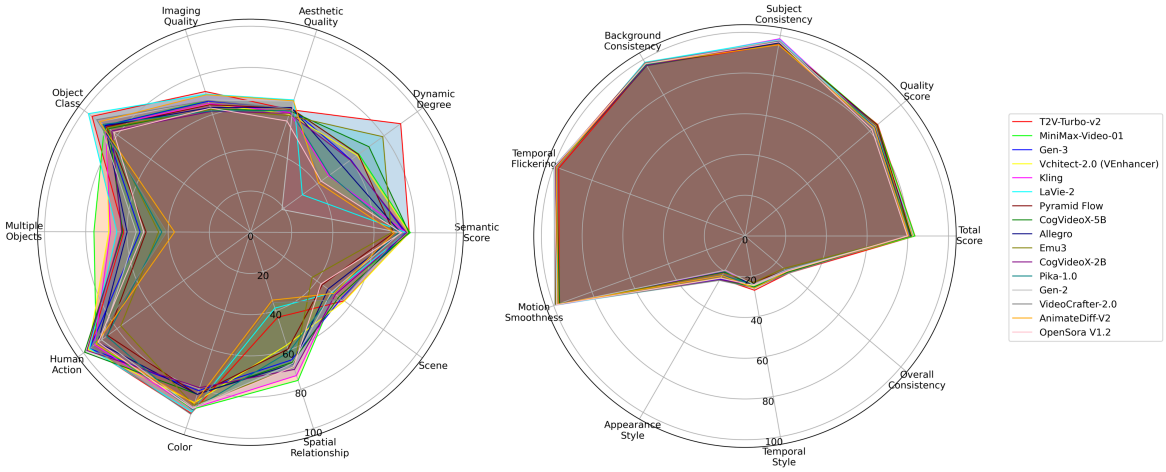

Fig. 8. Evaluation score of state-of-the-art models on VBench [177] - Left: High-variance metrics, Right: Low-variance metrics

TABLE VI
SUMMARY OF EVALUATION METRICS

Assessment target	Metric	Need ref. data	Need ref. model	What does it measure?
Spatial	↓FID	✓(Set)	×	similarity between the generated and original images' distribution
	↑PSNR	✓(Pair)	×	reconstruction quality
	↑IS	×	✓	image quality & diversity using a pretrained Inception network & conditional label distribution
	↑SSIM	✓(Pair)	×	brightness, contrast, and structural attributes
Temporal	↓FVD	✓(Set)	✓	similarity of feature space distribution of the generated video and the real ones
	↓KVD	✓(Set)	✓	appearance and motion
	↓FVDM	✓(Set)	✓	motion consistency
	↑DOVER	×	✓	aesthetic and technical perspectives
Text-Video Alignment	↑CLIP-score	×	✓	alignment of images and text using CLIP embedding
	↑BLIP	×	✓	feature space similarity between text and image
	↑viCLIP	×	✓	feature space similarity between text and generated video

ment with textual prompts. DEVIL [184] is an evaluation protocol that specifically targets the dynamics dimension in T2V models. It proposes a benchmark encompassing various dynamic complexity levels and evaluates videos through a set of dynamics scores, including dynamics range, controllability, and dynamics-based quality. The dynamics range measures the T2V model's capability to handle subtle and dramatic changes, while dynamic controllability assesses the model's ability to manipulate video dynamics based on text prompts. The dynamics-based quality metric evaluates the visual quality of generated content, via a composite quality score obtained by averaging the scores of various quality metrics correlated with dynamics.

V. DATASETS

Training datasets for the video generation task are notably scarce, especially in comparison to those available for image generation. In image generation, large-scale datasets like Laion-5B [48] are publicly available. By contrast, video generation models inherently require larger datasets and more computational resources for training, a demand that is especially pronounced in transformer-based models, which are known for their high data requirements. For instance, Movie Gen [35] training dataset consists of $O(100)$M video-text pairs and $O(1)$B image-text pairs. Similarly, Open-Sora [11] is trained on 30M videos, totaling roughly 80K hours of footage, and VideoPoet [70] utilizes a dataset of 1B image-text pairs in addition to approximately 270M videos.

Despite the scale, these datasets generally lack the detailed annotations necessary for more nuanced video generation tasks and often only describe scenes without capturing the camera motion or the temporal information crucial for understanding text-action dynamics in videos. Table VIII outlines some of the most widely used datasets currently available for video generation tasks.

VI. CHALLENGES AND FUTURE DIRECTIONS

Despite the unprecedented progress achieved in the field of video generation, several significant challenges persist. One major issue is *long duration generation*. Current video generation models are capable of producing plausible video clips ranging from 2 seconds to one minute. While some models have managed to generate videos extending several minutes, their spatial resolution, temporal consistency, and capacity for thematic generalization remain markedly limited. To create engaging video content, it is imperative to avoid the monotony associated with simple autoregressive extensions and instead develop compelling storylines and smooth semantic evolution. Notable challenges associated with extended durations include effective scene transitions, character development, and the

TABLE VII
COMPARISON OF STATE-OF-THE-ART MODELS ON VBENCH

Model Name	Total Score	Quality Score	Semantic Score	Motion Smoothness	Overall Consistency
T2V-Turbo-v2 [168]	**83.52**	**85.13**	77.12	97.07	**28.26**
MiniMax-Video-01 [62]	83.41	84.85	**77.65**	99.22	27.10
Gen-3 [56]	82.32	84.11	75.17	99.23	26.69
Vchitect-2.0 (VEnhancer) [125]	82.24	83.54	77.06	98.98	27.57
Kling [59]	81.85	83.39	75.68	99.40	26.42
LaVie-2 [15]	81.75	83.24	75.76	98.42	27.39
Pyramid Flow [12]	81.72	84.74	69.62	99.12	26.23
CogVideoX-5B [41]	81.61	82.75	77.04	96.92	27.59
Allegro [124]	81.09	83.12	72.98	98.82	26.36
Emu3 [185]	80.96	84.09	68.43	98.93	24.79
CogVideoX-2B [41]	80.91	82.18	75.83	97.73	26.66
Pika-1.0 [57]	80.69	82.92	71.77	99.50	25.94
Gen-2 [56]	80.58	82.47	73.03	**99.58**	26.17
VideoCrafter-2.0 [54]	80.44	82.20	73.42	97.73	28.23
AnimateDiff-V2 [37]	80.27	82.90	69.75	97.76	27.04
OpenSora1.2 [11]	79.76	81.35	73.39	98.50	26.85

TABLE VIII
OVERVIEW OF THE MOST COMMONLY USED DATASETS FOR VIDEO GENERATION TASK

Dataset	Year	#Clips	Clip Length(avg)	Total Duration(h)	Resolution	Modality
FineVideo [186]	2024	43.7K	4.7min	3.43K	diverse	video + text$^{\diamond}$
HowTo100M [187]	2019	136M	3.6s	134.5K	240p	video + text(instruction)
HD-VILA-100M [188]	2022	103M	13.4s	371.5K	720p	video + text
Webvid-10M [189]	2021	10M	18s	52K	360p	video + text
UCF-101 [190]	2012	13K	7s	27	240p	video + text(class label)
Kinect-600 [191]	2018	480K	10s	1.4K	-	video + text (action label)
MSR-VTT [192]	2016	10K	15s	41.2	240p	video + text
SkyTimelapse [193]	2018	35K	32 frames	-	640×360	video
Youtube-8M [194]	2016	8M	120-500s	350K	-	video + class label
X4K1000FPS [195]	2021	4.4K	65 frames	-	4096×2160	video w/ extreme motion(1000fps)
InternVid [103]	2023	234M	11.7s	760.3K	720p	video + text
Panda-70M [196]	2024	70.8M	8.5s	166.8K	720p	video + text
LSMDC [197]	2017	118K	4.8s	158	1080p	video + text
Youku-mPLUG [198]	2023	10M	54.2s	150K	-	video + text(Chinese)
VidGen-1M [199]	2024	1M	10.6s	-	720p	video + text
VidProM [200]	2024	6.7M	1.6-3s	4K	-	synthetic video + text
GenVideo [201]	2024	1M	2-6s	-	512 to 1280 px	synthetic video
COCO Caption [202]	2015	330K	-	-	640×480	image + text
LAION-5B [48]	2022	5.6B	-	-	256 to 1024 px	image + text

$^{\diamond}$ Detailed video description such as characters

enrichment of both action and plot.

To address these challenges, emerging methodologies involving multi-prompt/multi-scenario video generation [158], [159], [161] aim to compose coherent scenarios and maintain consistent personas. However, the content produced by these models often lacks aesthetic appeal and spatiotemporal quality. While entity-driven techniques have demonstrated promising results within the domain of image generation [132], similar progress in video generation remains elusive due to the additional constraints of maintaining temporal consistency. In both image and video generation, achieving fidelity and consistency in the representation of human entities is particularly challenging, as human observers exhibit heightened sensitivity to facial details compared to other objects.

High spatial and temporal resolution video generation poses further challenges in terms of quality and resource consumption. State-of-the-art closed-source video generation models have achieved resolutions of 1920×1080 @30 fps and open-source models reached $\approx 1280 \times 768$ @30 fps. Nevertheless, rendering sharp high-frequency details as depicted in Fig. 9-d, continues to be a challenge. Despite ongoing research on *flexible resolution and aspect ratio* vision transformers [203], [204], most existing video generation models do not yet provide this flexibility. Moreover, running inference for most state-of-the-art models requires cutting-edge GPUs with high VRAM capacity, which motivates research focused on *training and inference optimization*.

Consistency is another critical challenge in video generation, manifesting in various forms. *Temporal consistency* is especially vital when generating high-frequency details, such as leaves on trees as illustrated in Fig. 9-d. Similarly, *semantic-level consistency* remains an issue for many generative models. Often, the generated content fails to adhere to the rules of physics; for instance, solid objects may transform inappropriately during interactions with other objects - e.g., a brush melting into a canvas upon contact or chopsticks deforming when placed in the mouth as illustrated in Fig. 9-e. Besides, certain entities, such as human teeth or fingers, present particular difficulties in rendering, often resulting in deformed or inconsistent representations(Fig. 9-a and c). To mitigate these

issues, works such as GPT4Motion [205] propose leveraging large language models (LLMs) to generate Blender scripts based on user prompts. These scripts can then be used to create coherent motion and depth maps, aligning with physical rules through the Blender physics engine. However, this approach requires additional resources to run Blender and lacks scalability. Understanding and controlling *semantic-level composition* remains a challenge for most T2V generation models. As depicted in Fig. 9-b, these models frequently struggle to adhere to instructions regarding the semantic composition of scenes. While conditional video generation approaches aim to resolve these issues, providing additional input conditions in real-world applications can be limiting and not user-friendly, highlighting the interest of text-driven compositional video generation [161] as an open area of research.

Another major concern pertains to *dynamic level* and *motion control*. In general, T2V models tend to generate videos with low dynamism to achieve higher scores on quality metrics such as naturalness, motion smoothness, subject consistency, and background coherence [184]. Techniques for camera and motion control have been extensively explored in the literature, each presenting its advantages and disadvantages, as discussed in detail in Chapter III-C2. Notably, achieving sophisticated text-driven camera movements or coordinating multiple object motions while maintaining visual coherence remains an open area for development.

Multi-modal video generation represents an emerging research direction, paving the way toward Artificial General Intelligence (AGI) or human-level AI [185]. Vision-language models (VLMs) have demonstrated impressive performance in challenging tasks such as image/video captioning and visual question answering, thereby opening new avenues for research and enhancing existing applications. For instance, VLMs have been utilized to enrich video training data with enhanced captions. Joint audio-video generation has also garnered attention, expanding from audio-driven video generation to synthesizing audio that is synchronized with visual content.

Finally, the rapid evolution of video generation technology raises critical concerns regarding *security* and *ethical* implications. Generated videos may encompass illegal or unethical content, misinformation, or misattribution, and there is currently a lack of comprehensive quantitative understanding regarding their safety, thereby posing challenges to their reliability and practical deployment. It is essential to implement measures at multiple levels, including prompt filtering to prevent the generation of inappropriate content and post-generation filtering to detect and block unsafe videos. However, these safety measures are primarily applicable to service providers and are not enforceable on open-source models that could potentially be exploited by malicious individuals. Consequently, researchers have proposed techniques for embedding signatures within AI-generated content [206], [207], which are imperceptible by humans but critical for detection and consequently foster trust in the information.

REFERENCES

[1] T. M. Research, "Media (Video) Processing Solutions Market Growth, 2023-2031," 2023. [Online]. Available: https://www.transparencymarketresearch.com/media-processing-solutions-market.html

[2] "DALL·E 2." [Online]. Available: https://labs.openai.com

[3] C.-L. Cheang, G. Chen, Y. Jing, T. Kong, H. Li, Y. Li, Y. Liu, H. Wu, J. Xu, Y. Yang *et al.*, "Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation," *arXiv preprint arXiv:2410.06158*, 2024.

[4] H. Wu, Y. Jing, C. Cheang, G. Chen, J. Xu, X. Li, M. Liu, H. Li, and T. Kong, "Unleashing large-scale video generative pre-training for visual robot manipulation," *arXiv preprint arXiv:2312.13139*, 2023.

[5] J. Wu, H. Ma, C. Deng, and M. Long, "Pre-training contextualized world models with in-the-wild videos for reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

[7] H. Ni, B. Egger, S. Lohit, A. Cherian, Y. Wang, T. Koike-Akino, S. X. Huang, and T. K. Marks, "TI2V-Zero: Zero-Shot Image Conditioning for Text-to-Video Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9015–9025. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Ni_TI2V-Zero_Zero-Shot_Image_Conditioning_for_Text-to-Video_Diffusion_Models_CVPR_2024_paper.html

[8] Y. Wang, J. Bao, W. Weng, R. Feng, D. Yin, T. Yang, J. Zhang, Q. Dai, Z. Zhao, and C. Wang, "Microcinema: A divide-and-conquer approach for text-to-video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8414–8424. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Wang_MicroCinema_A_Divide-and-Conquer_Approach_for_Text-to-Video_Generation_CVPR_2024_paper.html

[9] C. Li, D. Huang, Z. Lu, Y. Xiao, Q. Pei, and L. Bai, "A Survey on Long Video Generation: Challenges, Methods, and Prospects," Mar. 2024, arXiv:2403.16407 [cs]. [Online]. Available: http://arxiv.org/abs/2403.16407

[10] A. Eassa and B. E. Sukru, "NVIDIA Sets New Generative AI Performance and Scale Records in MLPerf Training v4.0," Jun. 2024. [Online]. Available: https://developer.nvidia.com/blog/nvidia-sets-new-generative-ai-performance\-and-scale-records-in-mlperf-training-v4-0/

[11] Z. Zheng, X. Peng, T. Yang, C. Shen, S. Li, H. Liu, Y. Zhou, T. Li, and Y. You, "Open-sora: Democratizing efficient video production for all," March 2024. [Online]. Available: https://github.com/hpcaitech/Open-Sora

[12] Y. Jin, Z. Sun, N. Li, K. Xu, H. Jiang, N. Zhuang, Q. Huang, Y. Song, Y. Mu, and Z. Lin, "Pyramidal flow matching for efficient video generative modeling," *arXiv preprint arXiv:2410.05954*, 2024.

[13] X. Guo, J. Liu, M. Cui, and D. Huang, "I4VGen: Image as Stepping Stone for Text-to-Video Generation," Jun. 2024, arXiv:2406.02230 [cs]. [Online]. Available: http://arxiv.org/abs/2406.02230

[14] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada, A. Ephrat, J. Hur, G. Liu, A. Raj, Y. Li, M. Rubinstein, T. Michaeli, O. Wang, D. Sun, T. Dekel, and I. Mosseri, "Lumiere: A Space-Time Diffusion Model for Video Generation," Feb. 2024, arXiv:2401.12945 [cs]. [Online]. Available: http://arxiv.org/abs/2401.12945

[15] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, and P. Yang, "Lavie: High-quality video generation with cascaded latent diffusion models," *arXiv preprint arXiv:2309.15103*, 2023. [Online]. Available: https://arxiv.org/abs/2309.15103

[16] S. Yin, C. Wu, H. Yang, J. Wang, X. Wang, M. Ni, Z. Yang, L. Li, S. Liu, F. Yang, J. Fu, G. Ming, L. Wang, Z. Liu, H. Li, and N. Duan, "NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation," Mar. 2023, arXiv:2303.12346 [cs]. [Online]. Available: http://arxiv.org/abs/2303.12346

[17] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, and T. Salimans, "Imagen Video: High Definition Video Generation with Diffusion Models," Oct. 2022, arXiv:2210.02303 [cs]. [Online]. Available: http://arxiv.org/abs/2210.02303

[18] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers," May 2022, arXiv:2205.15868 [cs]. [Online]. Available: http://arxiv.org/abs/2205.15868

[19] J. Liang, C. Wu, X. Hu, Z. Gan, J. Wang, L. Wang, Z. Liu, Y. Fang, and N. Duan, "Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis," *Advances in Neural Information Processing Systems*, vol. 35, pp. 15 420–15 432, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/6358cd0cd6607fdf4870595795eb1710-Abstract-Conference.html

[20] V. Voleti, A. Jolicoeur-Martineau, and C. Pal, "Mcvd-masked conditional video diffusion for prediction, generation, and interpolation," *Advances in neural information processing systems*, vol. 35, pp. 23 371–23 385, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/944618542d80a63bbec16dfbd2bd689a-Abstract-Conference.html

[21] I. Skorokhodov, S. Tulyakov, and M. Elhoseiny, "Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3626–3636. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2022/html/Skorokhodov_StyleGAN-V_A_Continuous_Video_Generator_With_the_Price_Image_Quality_CVPR_2022_paper.html

[22] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman, "Make-A-Video: Text-to-Video Generation without Text-Video Data," Sep. 2022, arXiv:2209.14792 [cs]. [Online]. Available: http://arxiv.org/abs/2209.14792

[23] D. Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv: 2010.11929*, 2020.

[24] "Video generation models as world simulators | OpenAI." [Online]. Available: https://openai.com/index/video-generation-models-as-world-simulators/

[25] X. Ma, Y. Wang, G. Jia, X. Chen, Z. Liu, Y.-F. Li, C. Chen, and Y. Qiao, "Latte: Latent Diffusion Transformer for Video Generation," Jan. 2024, arXiv:2401.03048 [cs]. [Online]. Available: http://arxiv.org/abs/2401.03048

[26] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[27] A. Van Den Oord and O. Vinyals, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html

[28] L. Yu, Y. Cheng, K. Sohn, J. Lezama, H. Zhang, H. Chang, A. G. Hauptmann, M.-H. Yang, Y. Hao, and I. Essa, "Magvit: Masked generative video transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 459–10 469. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2023/html/Yu_MAGVIT_Masked_Generative_Video_Transformer_CVPR_2023_paper.html

[29] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, V. Birodkar, A. Gupta, X. Gu, A. G. Hauptmann, B. Gong, M.-H. Yang, I. Essa, D. A. Ross, and L. Jiang, "Language Model Beats Diffusion – Tokenizer is Key to Visual Generation," Mar. 2024, arXiv:2310.05737 [cs]. [Online]. Available: http://arxiv.org/abs/2310.05737

[30] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html

[31] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1526–1535.

[32] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *Advances in neural information processing systems*, vol. 31, 2018.

[33] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[34] M. Kumar, M. Babaeizadeh, D. Erhan, C. Finn, S. Levine, L. Dinh, and D. Kingma, "Videoflow: A conditional flow-based model for stochastic video generation," *arXiv preprint arXiv:1903.01434*, 2019. [Online]. Available: https://arxiv.org/abs/1903.01434

[35] "How Meta Movie Gen could usher in a new AI-enabled era for content creators." [Online]. Available: https://ai.meta.com/blog/movie-gen-media-foundation-models-generative-ai-video/

[36] Z. Duan, W. Zhou, C. Chen, Y. Li, and W. Qian, "ExVideo: Extending Video Diffusion Models via Parameter-Efficient Post-Tuning," Jun. 2024, arXiv:2406.14130 [cs]. [Online]. Available: http://arxiv.org/abs/2406.14130

[37] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, "AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning," Feb. 2024, arXiv:2307.04725 [cs]. [Online]. Available: http://arxiv.org/abs/2307.04725

[38] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 563–22 575. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2023/html/Blattmann_Align_Your_Latents_High-Resolution_Video_Synthesis_With_Latent_Diffusion_Models_CVPR_2023_paper.html

[39] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/39235c56aef13fb05a6adc95eb9d8d66-Abstract-Conference.html

[40] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial Attention in Multidimensional Transformers," Dec. 2019, arXiv:1912.12180 [cs]. [Online]. Available: http://arxiv.org/abs/1912.12180

[41] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, D. Yin, X. Gu, Y. Zhang, W. Wang, Y. Cheng, T. Liu, B. Xu, Y. Dong, and J. Tang, "CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer," Aug. 2024, arXiv:2408.06072 [cs]. [Online]. Available: http://arxiv.org/abs/2408.06072

[42] C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang, and N. Duan, "NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, vol. 13676, pp. 720–736, series Title: Lecture Notes in Computer Science. [Online]. Available: https://link.springer.com/10.1007/978-3-031-19787-1_41

[43] C. Wu, L. Huang, Q. Zhang, B. Li, L. Ji, F. Yang, G. Sapiro, and N. Duan, "Godiva: Generating open-domain videos from natural descriptions," *arXiv preprint arXiv:2104.14806*, 2021. [Online]. Available: https://arxiv.org/abs/2104.14806

[44] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[45] V. T. Hu, S. A. Baumann, M. Gui, O. Grebenkova, P. Ma, J. Fischer, and B. Ommer, "Zigma: Zigzag mamba diffusion model," *arXiv preprint arXiv:2403.13802*, 2024.

[46] Y. Gao, J. Huang, X. Sun, Z. Jie, Y. Zhong, and L. Ma, "Matten: Video generation with mamba-attention," *arXiv preprint arXiv:2405.03025*, 2024.

[47] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7346–7356. [Online]. Available: http://openaccess.thecvf.com/content/ICCV2023/html/Esser_Structure_and_Content-Guided_Video_Synthesis_with_Diffusion_Models_ICCV_2023_paper.html

[48] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, and M. Wortsman, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html

[49] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, and D. Erhan, "Phenaki: Variable length video generation from open domain textual descriptions," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=vOEXS39nOF

[50] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 954–15 964. [Online]. Available: http://openaccess.thecvf.com/content/ICCV2023/html/Khachatryan_

Text2Video-Zero_Text-to-Image_Diffusion_Models_are_Zero-Shot_Video_Generators_ICCV_2023_paper.html
[51] R. Girdhar, M. Singh, A. Brown, Q. Duval, S. Azadi, S. S. Rambhatla, A. Shah, X. Yin, D. Parikh, and I. Misra, "Emu Video: Factorizing Text-to-Video Generation by Explicit Image Conditioning," Aug. 2024, arXiv:2311.10709 [cs]. [Online]. Available: http://arxiv.org/abs/2311.10709
[52] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets," Nov. 2023, arXiv:2311.15127 [cs]. [Online]. Available: http://arxiv.org/abs/2311.15127
[53] H. Chen, M. Xia, Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang, C. Weng, and Y. Shan, "VideoCrafter1: Open Diffusion Models for High-Quality Video Generation," Oct. 2023, arXiv:2310.19512 [cs]. [Online]. Available: http://arxiv.org/abs/2310.19512
[54] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7310–7320. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Chen_VideoCrafter2_Overcoming_Data_Limitations_for_High-Quality_Video_Diffusion_Models_CVPR_2024_paper.html
[55] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, "ModelScope Text-to-Video Technical Report," Aug. 2023, arXiv:2308.06571 [cs]. [Online]. Available: http://arxiv.org/abs/2308.06571
[56] "Runway Research | Gen-2: Generate novel videos with text, images or video clips." [Online]. Available: https://runwayml.com/research/gen-2
[57] "Pika." [Online]. Available: https://pika.art/home
[58] "Haiper - AI Video Generator." [Online]. Available: https://haiper.ai/
[59] "KLING." [Online]. Available: https://kling.kuaishou.com/en
[60] F. Bao, C. Xiang, G. Yue, G. He, H. Zhu, K. Zheng, M. Zhao, S. Liu, Y. Wang, and J. Zhu, "Vidu: a Highly Consistent, Dynamic and Skilled Text-to-Video Generator with Diffusion Models," May 2024, arXiv:2405.04233 [cs]. [Online]. Available: http://arxiv.org/abs/2405.04233
[61] "Veo," Aug. 2024. [Online]. Available: https://deepmind.google/technologies/veo/
[62] "MiniMax." [Online]. Available: https://platform.minimaxi.com/
[63] V. Voleti, C.-H. Yao, M. Boss, A. Letts, D. Pankratz, D. Tochilkin, C. Laforte, R. Rombach, and V. Jampani, "Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion," *arXiv preprint arXiv:2403.12008*, 2024. [Online]. Available: https://arxiv.org/abs/2403.12008
[64] S. Zhang, J. Wang, Y. Zhang, K. Zhao, H. Yuan, Z. Qin, X. Wang, D. Zhao, and J. Zhou, "I2VGen-XL: High-Quality Image-to-Video Synthesis via Cascaded Diffusion Models," Nov. 2023, arXiv:2311.04145 [cs]. [Online]. Available: http://arxiv.org/abs/2311.04145
[65] J. Xing, M. Xia, Y. Zhang, H. Chen, W. Yu, H. Liu, X. Wang, T.-T. Wong, and Y. Shan, "DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors," Nov. 2023, arXiv:2310.12190 [cs]. [Online]. Available: http://arxiv.org/abs/2310.12190
[66] Y. Zhang, Z. Xing, Y. Zeng, Y. Fang, and K. Chen, "Pia: Your personalized image animator via plug-and-play modules in text-to-image models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7747–7756. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Zhang_PIA_Your_Personalized_Image_Animator_via_Plug-and-Play_Modules_in_Text-to-Image_CVPR_2024_paper.html
[67] X. Guo, M. Zheng, L. Hou, Y. Gao, Y. Deng, P. Wan, D. Zhang, Y. Liu, W. Hu, Z. Zha, H. Huang, and C. Ma, "I2V-Adapter: A General Image-to-Video Adapter for Diffusion Models," Jun. 2024, arXiv:2312.16693 [cs]. [Online]. Available: http://arxiv.org/abs/2312.16693
[68] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models," Aug. 2023, arXiv:2308.06721 [cs]. [Online]. Available: http://arxiv.org/abs/2308.06721
[69] P. Gao, L. Zhuo, D. Liu, R. Du, X. Luo, L. Qiu, Y. Zhang, C. Lin, R. Huang, S. Geng, R. Zhang, J. Xi, W. Shao, Z. Jiang, T. Yang, W. Ye, H. Tong, J. He, Y. Qiao, and H. Li, "Lumina-T2X: Transforming Text into Any Modality, Resolution, and Duration via Flow-based Large Diffusion Transformers," Jun. 2024, arXiv:2405.05945 [cs]. [Online]. Available: http://arxiv.org/abs/2405.05945
[70] D. Kondratyuk, L. Yu, X. Gu, J. Lezama, J. Huang, G. Schindler, R. Hornung, V. Birodkar, J. Yan, M.-C. Chiu, K. Somandepalli, H. Akbari, Y. Alon, Y. Cheng, J. Dillon, A. Gupta, M. Hahn, A. Hauth, D. Hendon, A. Martinez, D. Minnen, M. Sirotenko, K. Sohn, X. Yang, H. Adam, M.-H. Yang, I. Essa, H. Wang, D. A. Ross, B. Seybold, and L. Jiang, "VideoPoet: A Large Language Model for Zero-Shot Video Generation," Jun. 2024, arXiv:2312.14125 [cs]. [Online]. Available: http://arxiv.org/abs/2312.14125
[71] F. Reda, J. Kontkanen, E. Tabellion, D. Sun, C. Pantofaru, and B. Curless, "Film: Frame interpolation for large motion," in *European Conference on Computer Vision*. Springer, 2022, pp. 250–266. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-20071-7_15
[72] X. Chen, Y. Wang, L. Zhang, S. Zhuang, X. Ma, J. Yu, Y. Wang, D. Lin, Y. Qiao, and Z. Liu, "Seine: Short-to-long video diffusion model for generative transition and prediction," in *The Twelfth International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=FNq3nIvP4F
[73] J. Xing, H. Liu, M. Xia, Y. Zhang, X. Wang, Y. Shan, and T.-T. Wong, "ToonCrafter: Generative Cartoon Interpolation," May 2024, arXiv:2405.17933 [cs]. [Online]. Available: http://arxiv.org/abs/2405.17933
[74] Y. Guo, C. Yang, A. Rao, M. Agrawala, D. Lin, and B. Dai, "SparseCtrl: Adding Sparse Controls to Text-to-Video Diffusion Models," Nov. 2023, arXiv:2311.16933 [cs]. [Online]. Available: http://arxiv.org/abs/2311.16933
[75] Y. Zeng, G. Wei, J. Zheng, J. Zou, Y. Wei, Y. Zhang, and H. Li, "Make pixels dance: High-dynamic video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8850–8860. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Zeng_Make_Pixels_Dance_High-Dynamic_Video_Generation_CVPR_2024_paper.html
[76] S. Jain, D. Watson, E. Tabellion, B. Poole, and J. Kontkanen, "Video interpolation with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7341–7351. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Jain_Video_Interpolation_with_Diffusion_Models_CVPR_2024_paper.html
[77] Q. Wang, W. Li, C. Mou, X. Cheng, and J. Zhang, "360dvd: Controllable panorama video generation with 360-degree video diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6913–6923. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Wang_360DVD_Controllable_Panorama_Video_Generation_with_360-Degree_Video_Diffusion_Model_CVPR_2024_paper.html
[78] S. Yuan, J. Huang, Y. Shi, Y. Xu, R. Zhu, B. Lin, X. Cheng, L. Yuan, and J. Luo, "MagicTime: Time-lapse Video Generation Models as Metamorphic Simulators," Apr. 2024, arXiv:2404.05014 [cs]. [Online]. Available: http://arxiv.org/abs/2404.05014
[79] L. Liu, Q. Liu, S. Qian, Y. Zhou, W. Zhou, H. Li, L. Xie, and Q. Tian, "Text-Animator: Controllable Visual Text Video Generation," *arXiv preprint arXiv:2406.17777*, 2024. [Online]. Available: https://arxiv.org/abs/2406.17777
[80] L. Hu, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8153–8163. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Hu_Animate_Anyone_Consistent_and_Controllable_Image-to-Video_Synthesis_for_Character_Animation_CVPR_2024_paper.html
[81] R. Shao, Y. Pang, Z. Zheng, J. Sun, and Y. Liu, "Human4DiT: Free-view Human Video Generation with 4D Diffusion Transformer," May 2024, arXiv:2405.17405 [cs]. [Online]. Available: http://arxiv.org/abs/2405.17405
[82] X. Wang, S. Zhang, C. Gao, J. Wang, X. Zhou, Y. Zhang, L. Yan, and N. Sang, "UniAnimate: Taming Unified Video Diffusion Models for Consistent Human Image Animation," Jun. 2024, arXiv:2406.01188 [cs]. [Online]. Available: http://arxiv.org/abs/2406.01188
[83] Y. Ma, Y. He, X. Cun, X. Wang, S. Chen, X. Li, and Q. Chen, "Follow your pose: Pose-guided text-to-video generation using pose-free videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 4117–4125, issue: 5. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/28206
[84] Y. Xu, Y. Chen, Z. Huang, Z. He, G. Wang, P. Torr, and L. Lin, "AnimateZoo: Zero-shot Video Generation of Cross-Species Animation via Subject Alignment," Apr. 2024, arXiv:2404.04946 [cs]. [Online]. Available: http://arxiv.org/abs/2404.04946

[85] B. Qin, W. Ye, Q. Yu, S. Tang, and Y. Zhuang, "Dancing Avatar: Pose and Text-Guided Human Motion Videos Synthesis with Image Diffusion Model," Aug. 2023, arXiv:2308.07749 [cs]. [Online]. Available: http://arxiv.org/abs/2308.07749

[86] T. Wang, L. Li, K. Lin, Y. Zhai, C.-C. Lin, Z. Yang, H. Zhang, Z. Liu, and L. Wang, "Disco: Disentangled control for realistic human dance generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9326–9336. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Wang_DisCo_Disentangled_Control_for_Realistic_Human_Dance_Generation_CVPR_2024_paper.html

[87] Z. Xu, J. Zhang, J. H. Liew, H. Yan, J.-W. Liu, C. Zhang, J. Feng, and M. Z. Shou, "Magicanimate: Temporally consistent human image animation using diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1481–1490. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Xu_MagicAnimate_Temporally_Consistent_Human_Image_Animation_using_Diffusion_Model_CVPR_2024_paper.html

[88] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman, "Dreampose: Fashion image-to-video synthesis via stable diffusion," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 22 623–22 633. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10377471/

[89] H. Wei, Z. Yang, and Z. Wang, "AniPortrait: Audio-Driven Synthesis of Photorealistic Portrait Animation," Mar. 2024, arXiv:2403.17694 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2403.17694

[90] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8652–8661. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2023/html/Zhang_SadTalker_Learning_Realistic_3D_Motion_Coefficients_for_Stylized_Audio-Driven_Single_CVPR_2023_paper.html

[91] Z. Chen, J. Cao, Z. Chen, Y. Li, and C. Ma, "EchoMimic: Lifelike Audio-Driven Portrait Animations through Editable Landmark Conditions," Jul. 2024, arXiv:2407.08136 [cs]. [Online]. Available: http://arxiv.org/abs/2407.08136

[92] J. Xing, M. Xia, Y. Zhang, X. Cun, J. Wang, and T.-T. Wong, "Codetalker: Speech-driven 3d facial animation with discrete motion prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 780–12 790. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2023/html/Xing_CodeTalker_Speech-Driven_3D_Facial_Animation_With_Discrete_Motion_Prior_CVPR_2023_paper.html

[93] L. Tian, Q. Wang, B. Zhang, and L. Bo, "EMO: Emote Portrait Alive – Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions," Aug. 2024, arXiv:2402.17485 [cs]. [Online]. Available: http://arxiv.org/abs/2402.17485

[94] Q. He, X. Ji, Y. Gong, Y. Lu, Z. Diao, L. Huang, Y. Yao, S. Zhu, Z. Ma, S. Xu, X. Wu, Z. Zhang, X. Cao, and H. Zhu, "EmoTalk3D: High-Fidelity Free-View Synthesis of Emotional 3D Talking Head," Aug. 2024, arXiv:2408.00297 [cs]. [Online]. Available: http://arxiv.org/abs/2408.00297

[95] M. Xu, H. Li, Q. Su, H. Shang, L. Zhang, C. Liu, J. Wang, Y. Yao, and S. Zhu, "Hallo: Hierarchical Audio-Driven Visual Synthesis for Portrait Image Animation," Jun. 2024, arXiv:2406.08801 [cs]. [Online]. Available: http://arxiv.org/abs/2406.08801

[96] G. Kim, H. Shim, H. Kim, Y. Choi, J. Kim, and E. Yang, "Diffusion video autoencoders: Toward temporally consistent face video editing via disentangled video encoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6091–6100. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2023/html/Kim_Diffusion_Video_Autoencoders_Toward_Temporally_Consistent_Face_Video_Editing_via_CVPR_2023_paper.html

[97] X. He, Q. Liu, S. Qian, X. Wang, T. Hu, K. Cao, K. Yan, and J. Zhang, "ID-Animator: Zero-Shot Identity-Preserving Human Video Generation," Jun. 2024, arXiv:2404.15275 [cs]. [Online]. Available: http://arxiv.org/abs/2404.15275

[98] Y. Wang, J. Guo, J. Bai, R. Yu, T. He, X. Tan, X. Sun, and J. Bian, "InstructAvatar: Text-Guided Emotion and Motion Control for Avatar Generation," May 2024, arXiv:2405.15758 [cs]. [Online]. Available: http://arxiv.org/abs/2405.15758

[99] J. Guo, D. Zhang, X. Liu, Z. Zhong, Y. Zhang, P. Wan, and D. Zhang, "LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control," Jul. 2024, arXiv:2407.03168 [cs]. [Online]. Available: http://arxiv.org/abs/2407.03168

[100] Y. Ma, S. Wang, Z. Hu, C. Fan, T. Lv, Y. Ding, Z. Deng, and X. Yu, "Styletalk: One-shot talking head generation with controllable speaking styles," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1896–1904.

[101] H. Yuan, S. Zhang, X. Wang, Y. Wei, T. Feng, Y. Pan, Y. Zhang, Z. Liu, S. Albanie, and D. Ni, "InstructVideo: instructing video diffusion models with human feedback," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6463–6474. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Yuan_InstructVideo_Instructing_Video_Diffusion_Models_with_Human_Feedback_CVPR_2024_paper.html

[102] M. Prabhudesai, R. Mendonca, Z. Qin, K. Fragkiadaki, and D. Pathak, "Video Diffusion Alignment via Reward Gradients," Jul. 2024, arXiv:2407.08737 [cs]. [Online]. Available: http://arxiv.org/abs/2407.08737

[103] Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Li, G. Chen, X. Chen, Y. Wang, C. He, P. Luo, Z. Liu, Y. Wang, L. Wang, and Y. Qiao, "InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation," Jan. 2024, arXiv:2307.06942 [cs]. [Online]. Available: http://arxiv.org/abs/2307.06942

[104] Z. Tang, Z. Yang, M. Khademi, Y. Liu, C. Zhu, and M. Bansal, "CoDi-2: In-Context Interleaved and Interactive Any-to-Any Generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 425–27 434. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2024/html/Tang_CoDi-2_In-Context_Interleaved_and_Interactive_Any-to-Any_Generation_CVPR_2024_paper.html

[105] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models," Jun. 2024, arXiv:2306.05424 [cs]. [Online]. Available: http://arxiv.org/abs/2306.05424

[106] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, vol. 1, no. 2, p. 3, 2023.

[107] M. Ku, C. Wei, W. Ren, H. Yang, and W. Chen, "AnyV2V: A Tuning-Free Framework For Any Video-to-Video Editing Tasks," Jun. 2024, arXiv:2403.14468 [cs]. [Online]. Available: http://arxiv.org/abs/2403.14468

[108] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7623–7633. [Online]. Available: http://openaccess.thecvf.com/content/ICCV2023/html/Wu_Tune-A-Video_One-Shot_Tuning_of_Image_Diffusion_Models_for_Text-to-Video_Generation_ICCV_2023_paper.html

[109] S. Liu, Y. Zhang, W. Li, Z. Lin, and J. Jia, "Video-p2p: Video editing with cross-attention control," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8599–8608. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Video-P2P_Video_Editing_with_Cross-attention_Control_CVPR_2024_paper.html

[110] J. Bai, T. He, Y. Wang, J. Guo, H. Hu, Z. Liu, and J. Bian, "UniEdit: A Unified Tuning-Free Framework for Video Motion and Appearance Editing," Apr. 2024, arXiv:2402.13185 [cs]. [Online]. Available: http://arxiv.org/abs/2402.13185

[111] H. Ouyang, Q. Wang, Y. Xiao, Q. Bai, J. Zhang, K. Zheng, X. Zhou, Q. Chen, and Y. Shen, "Codef: Content deformation fields for temporally consistent video processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8089–8099. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Ouyang_CoDeF_Content_Deformation_Fields_for_Temporally_Consistent_Video_Processing_CVPR_2024_paper.html

[112] M. Geyer, O. Bar-Tal, S. Bagon, and T. Dekel, "TokenFlow: Consistent Diffusion Features for Consistent Video Editing," Nov. 2023, arXiv:2307.10373 [cs]. [Online]. Available: http://arxiv.org/abs/2307.10373

[113] J. Cheng, T. Xiao, and T. He, "Consistent Video-to-Video Transfer Using Synthetic Dataset," Dec. 2023, arXiv:2311.00213 [cs]. [Online]. Available: http://arxiv.org/abs/2311.00213

[114] C. Qi, X. Cun, Y. Zhang, C. Lei, X. Wang, Y. Shan, and Q. Chen, "Fatezero: Fusing attentions for zero-shot text-based video editing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 932–15 942. [Online]. Available:

http://openaccess.thecvf.com/content/ICCV2023/html/QI_FateZero_Fusing_Attentions_for_Zero-shot_Text-based_Video_Editing_ICCV_2023_paper.html

[115] C. Mou, M. Cao, X. Wang, Z. Zhang, Y. Shan, and J. Zhang, "ReVideo: Remake a Video with Motion and Content Control," May 2024, arXiv:2405.13865 [cs]. [Online]. Available: http://arxiv.org/abs/2405.13865

[116] W. Chai, X. Guo, G. Wang, and Y. Lu, "Stablevideo: Text-driven consistency-aware diffusion video editing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 040–23 050. [Online]. Available: http://openaccess.thecvf.com/content/ICCV2023/html/Chai_StableVideo_Text-driven_Consistency-aware_Diffusion_Video_Editing_ICCV_2023_paper.html

[117] W. Chen, Y. Ji, J. Wu, H. Wu, P. Xie, J. Li, X. Xia, X. Xiao, and L. Lin, "Control-A-Video: Controllable Text-to-Video Diffusion Models with Motion Prior and Reward Feedback Learning," Aug. 2024, arXiv:2305.13840 [cs]. [Online]. Available: http://arxiv.org/abs/2305.13840

[118] O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel, "Text2LIVE: Text-Driven Layered Image and Video Editing," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, vol. 13675, pp. 707–723, series Title: Lecture Notes in Computer Science. [Online]. Available: https://link.springer.com/10.1007/978-3-031-19784-0_41

[119] D. Ceylan, C.-H. P. Huang, and N. J. Mitra, "Pix2video: Video editing using image diffusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 206–23 217. [Online]. Available: http://openaccess.thecvf.com/content/ICCV2023/html/Ceylan_Pix2Video_Video_Editing_using_Image_Diffusion_ICCV_2023_paper.html

[120] G. Liu, M. Xia, Y. Zhang, H. Chen, J. Xing, X. Wang, Y. Yang, and Y. Shan, "StyleCrafter: Enhancing Stylized Text-to-Video Generation with Style Adapter," Nov. 2023, arXiv:2312.00330 [cs]. [Online]. Available: http://arxiv.org/abs/2312.00330

[121] Y. Xie, C.-H. Yao, V. Voleti, H. Jiang, and V. Jampani, "SV4D: Dynamic 3D Content Generation with Multi-Frame and Multi-View Consistency," Jul. 2024, arXiv:2407.17470 [cs]. [Online]. Available: http://arxiv.org/abs/2407.17470

[122] Team, "Luma launches Dream Machine." [Online]. Available: https://blog.lumalabs.ai/p/dream-machine

[123] [Online]. Available: https://www.genmo.ai/blog

[124] Y. Zhou, Q. Wang, Y. Cai, and H. Yang, "Allegro: Open the black box of commercial-level video generation model," *arXiv preprint arXiv:2410.15458*, 2024.

[125] J. He, T. Xue, D. Liu, X. Lin, P. Gao, D. Lin, Y. Qiao, W. Ouyang, and Z. Liu, "VEnhancer: Generative Space-Time Enhancement for Video Generation," Jul. 2024, arXiv:2407.07667 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2407.07667

[126] Y. Xu, T. Park, R. Zhang, Y. Zhou, E. Shechtman, F. Liu, J.-B. Huang, and D. Liu, "VideoGigaGAN: Towards Detail-rich Video Super-Resolution," May 2024, arXiv:2404.12388 [cs]. [Online]. Available: http://arxiv.org/abs/2404.12388

[127] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "Basicvsr++: Improving video super-resolution with enhanced propagation and alignment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5972–5981. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2022/html/Chan_BasicVSR_Improving_Video_Super-Resolution_With_Enhanced_Propagation_and_Alignment_CVPR_2022_paper.html

[128] H. Ni, C. Shi, K. Li, S. X. Huang, and M. R. Min, "Conditional image-to-video generation with latent flow diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18 444–18 455. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2023/html/Ni_Conditional_Image-to-Video_Generation_With_Latent_Flow_Diffusion_Models_CVPR_2023_paper.html

[129] Y. Zhang, Y. Wei, X. Lin, Z. Hui, P. Ren, X. Xie, X. Ji, and W. Zuo, "VideoElevator: Elevating Video Generation Quality with Versatile Text-to-Image Diffusion Models," Mar. 2024, arXiv:2403.05438 [cs]. [Online]. Available: http://arxiv.org/abs/2403.05438

[130] T. Wu, C. Si, Y. Jiang, Z. Huang, and Z. Liu, "FreeInit: Bridging Initialization Gap in Video Diffusion Models," Jul. 2024, arXiv:2312.07537 [cs]. [Online]. Available: http://arxiv.org/abs/2312.07537

[131] X. Shi, Z. Huang, F.-Y. Wang, W. Bian, D. Li, Y. Zhang, M. Zhang, K. C. Cheung, S. See, H. Qin, J. Dai, and H. Li, "Motion-I2V: Consistent and Controllable Image-to-Video Generation with Explicit Motion Modeling," in *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24*. Denver CO USA: ACM, Jul. 2024, pp. 1–11. [Online]. Available: https://dl.acm.org/doi/10.1145/3641519.3657497

[132] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 500–22 510. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2023/html/Ruiz_DreamBooth_Fine_Tuning_Text-to-Image_Diffusion_Models_for_Subject-Driven_Generation_CVPR_2023_paper.html

[133] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, "Videocomposer: Compositional video synthesis with motion controllability," *Advances in Neural Information Processing Systems*, vol. 36, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/180f6184a3458fa19c28c5483bc61877-Abstract-Conference.html

[134] S. Yang, L. Hou, H. Huang, C. Ma, P. Wan, D. Zhang, X. Chen, and J. Liao, "Direct-a-Video: Customized Video Generation with User-Directed Camera Movement and Object Motion," in *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24*. Denver CO USA: ACM, Jul. 2024, pp. 1–12. [Online]. Available: https://dl.acm.org/doi/10.1145/3641519.3657481

[135] R. Zhao, Y. Gu, J. Z. Wu, D. J. Zhang, J. Liu, W. Wu, J. Keppo, and M. Z. Shou, "MotionDirector: Motion Customization of Text-to-Video Diffusion Models," Oct. 2023, arXiv:2310.08465 [cs]. [Online]. Available: http://arxiv.org/abs/2310.08465

[136] H. He, Y. Xu, Y. Guo, G. Wetzstein, B. Dai, H. Li, and C. Yang, "CameraCtrl: Enabling Camera Control for Text-to-Video Generation," Apr. 2024, arXiv:2404.02101. [Online]. Available: http://arxiv.org/abs/2404.02101

[137] C. Hou, G. Wei, Y. Zeng, and Z. Chen, "Training-free Camera Control for Video Generation," Jun. 2024, arXiv:2406.10126 [cs]. [Online]. Available: http://arxiv.org/abs/2406.10126

[138] M. Zhao, H. Zhu, C. Xiang, K. Zheng, C. Li, and J. Zhu, "Identifying and Solving Conditional Image Leakage in Image-to-Video Diffusion Model," Jun. 2024, arXiv:2406.15735 [cs]. [Online]. Available: http://arxiv.org/abs/2406.15735

[139] T.-S. Chen, C. H. Lin, H.-Y. Tseng, T.-Y. Lin, and M.-H. Yang, "Motion-Conditioned Diffusion Model for Controllable Video Synthesis," Apr. 2023, arXiv:2304.14404 [cs]. [Online]. Available: http://arxiv.org/abs/2304.14404

[140] W. Wu, Z. Li, Y. Gu, R. Zhao, Y. He, D. J. Zhang, M. Z. Shou, Y. Li, T. Gao, and D. Zhang, "DragAnything: Motion Control for Anything using Entity Representation," Mar. 2024, arXiv:2403.07420 [cs]. [Online]. Available: http://arxiv.org/abs/2403.07420

[141] S. Yin, C. Wu, J. Liang, J. Shi, H. Li, G. Ming, and N. Duan, "DragNUWA: Fine-grained Control in Video Generation by Integrating Text, Image, and Trajectory," Aug. 2023, arXiv:2308.08089 [cs]. [Online]. Available: http://arxiv.org/abs/2308.08089

[142] Z. Wang, Z. Yuan, X. Wang, Y. Li, T. Chen, M. Xia, P. Luo, and Y. Shan, "MotionCtrl: A Unified and Flexible Motion Controller for Video Generation," in *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24*. Denver CO USA: ACM, Jul. 2024, pp. 1–11. [Online]. Available: https://dl.acm.org/doi/10.1145/3641519.3657518

[143] J. Wang, Y. Zhang, J. Zou, Y. Zeng, G. Wei, L. Yuan, and H. Li, "Boximator: Generating Rich and Controllable Motions for Video Synthesis," Feb. 2024, arXiv:2402.01566 [cs]. [Online]. Available: http://arxiv.org/abs/2402.01566

[144] P. Ling, J. Bu, P. Zhang, X. Dong, Y. Zang, T. Wu, H. Chen, J. Wang, and Y. Jin, "MotionClone: Training-Free Motion Cloning for Controllable Video Generation," Jun. 2024, arXiv:2406.05338 [cs]. [Online]. Available: http://arxiv.org/abs/2406.05338

[145] D. Yatim, R. Fridman, O. Bar-Tal, Y. Kasten, and T. Dekel, "Space-time diffusion features for zero-shot text-driven motion transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8466–8476. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Yatim_Space-Time_Diffusion_Features_for_Zero-Shot_Text-Driven_Motion_Transfer_CVPR_2024_paper.html

[146] Y. Ren, Y. Zhou, J. Yang, J. Shi, D. Liu, F. Liu, M. Kwon, and A. Shrivastava, "Customize-A-Video: One-Shot Motion Customization of Text-to-Video Diffusion Models," Feb. 2024, arXiv:2402.14780 [cs]. [Online]. Available: http://arxiv.org/abs/2402.14780

[147] H. Jeong, G. Y. Park, and J. C. Ye, "Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9212–9221. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Jeong_VMC_Video_Motion_Customization_using_Temporal_Attention_Adaption_for_Text-to-Video_CVPR_2024_paper.html

[148] J. Wu, X. Li, Y. Zeng, J. Zhang, Q. Zhou, Y. Li, Y. Tong, and K. Chen, "MotionBooth: Motion-Aware Customized Text-to-Video Generation," Jun. 2024, arXiv:2406.17758 [cs]. [Online]. Available: http://arxiv.org/abs/2406.17758

[149] Y. Wei, S. Zhang, Z. Qing, H. Yuan, Z. Liu, Y. Liu, Y. Zhang, J. Zhou, and H. Shan, "Dreamvideo: Composing your dream videos with customized subject and motion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6537–6549. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Wei_DreamVideo_Composing_Your_Dream_Videos_with_Customized_Subject_and_Motion_CVPR_2024_paper.html

[150] Y. Jiang, T. Wu, S. Yang, C. Si, D. Lin, Y. Qiao, C. C. Loy, and Z. Liu, "Videobooth: Diffusion-based video generation with image prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6689–6700. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Jiang_VideoBooth_Diffusion-based_Video_Generation_with_Image_Prompts_CVPR_2024_paper.html

[151] R. Henschel, L. Khachatryan, D. Hayrapetyan, H. Poghosyan, V. Tadevosyan, Z. Wang, S. Navasardyan, and H. Shi, "StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text," Mar. 2024, arXiv:2403.14773 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2403.14773

[152] H. Qiu, M. Xia, Y. Zhang, Y. He, X. Wang, Y. Shan, and Z. Liu, "FreeNoise: Tuning-Free Longer Video Diffusion via Noise Rescheduling," Jan. 2024, arXiv:2310.15169 [cs]. [Online]. Available: http://arxiv.org/abs/2310.15169

[153] S. Zhuang, K. Li, X. Chen, Y. Wang, Z. Liu, Y. Qiao, and Y. Wang, "Vlogger: Make your dream a vlog," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8806–8817. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Zhuang_Vlogger_Make_Your_Dream_A_Vlog_CVPR_2024_paper.html

[154] F.-Y. Wang, W. Chen, G. Song, H.-J. Ye, Y. Liu, and H. Li, "Gen-L-Video: Multi-Text to Long Video Generation via Temporal Co-Denoising," May 2023, arXiv:2305.18264. [Online]. Available: http://arxiv.org/abs/2305.18264

[155] W. Harvey, S. Naderiparizi, V. Masrani, C. Weilbach, and F. Wood, "Flexible diffusion modeling of long videos," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 953–27 965, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/b2fe1ee8d936ac08dd26f2ff58986c8f-Abstract-Conference.html

[156] J. Kim, J. Kang, J. Choi, and B. Han, "FIFO-Diffusion: Generating Infinite Videos from Text without Training," Jun. 2024, arXiv:2405.11473 [cs]. [Online]. Available: http://arxiv.org/abs/2405.11473

[157] J. Gu, S. Wang, H. Zhao, T. Lu, X. Zhang, Z. Wu, S. Xu, W. Zhang, Y.-G. Jiang, and H. Xu, "Reuse and Diffuse: Iterative Denoising for Text-to-Video Generation," Sep. 2023, arXiv:2309.03549 [cs]. [Online]. Available: http://arxiv.org/abs/2309.03549

[158] F. Long, Z. Qiu, T. Yao, and T. Mei, "VideoDrafter: Content-Consistent Multi-Scene Video Generation with LLM," Jan. 2024, arXiv:2401.01256 [cs]. [Online]. Available: http://arxiv.org/abs/2401.01256

[159] H. Lin, A. Zala, J. Cho, and M. Bansal, "VideoDirectorGPT: Consistent Multi-scene Video Generation via LLM-Guided Planning," Jul. 2024, arXiv:2309.15091 [cs]. [Online]. Available: http://arxiv.org/abs/2309.15091

[160] G. Oh, J. Jeong, S. Kim, W. Byeon, J. Kim, S. Kim, and S. Kim, "MEVG: Multi-event Video Generation with Text-to-Video Models," Jul. 2024, arXiv:2312.04086 [cs]. [Online]. Available: http://arxiv.org/abs/2312.04086

[161] Y. Tian, L. Yang, H. Yang, Y. Gao, Y. Deng, J. Chen, X. Wang, Z. Yu, X. Tao, P. Wan, D. Zhang, and B. Cui, "VideoTetris: Towards Compositional Text-to-Video Generation," Jun. 2024, arXiv:2406.04277 [cs]. [Online]. Available: http://arxiv.org/abs/2406.04277

[162] Y. Zhao, L. Zhao, X. Zhou, J. Wu, C.-T. Chu, H. Miao, F. Schroff, H. Adam, T. Liu, B. Gong *et al.*, "Distilling vision-language models on millions of videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 106–13 116.

[163] X. Wang, S. Zhang, H. Yuan, Z. Qing, B. Gong, Y. Zhang, Y. Shen, C. Gao, and N. Sang, "A recipe for scaling up text-to-video generation with text-free videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6572–6582. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Wang_A_Recipe_for_Scaling_up_Text-to-Video_Generation_with_Text-free_Videos_CVPR_2024_paper.html

[164] T. Zhao, T. Fang, E. Liu, W. Rui, W. Soedarmadji, S. Li, Z. Lin, G. Dai, S. Yan, H. Yang *et al.*, "Vidit-q: Efficient and accurate quantization of diffusion transformers for image and video generation," *arXiv preprint arXiv:2406.02540*, 2024.

[165] X. Li, Y. Liu, L. Lian, H. Yang, Z. Dong, D. Kang, S. Zhang, and K. Keutzer, "Q-diffusion: Quantizing diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 535–17 545.

[166] Y. He, L. Liu, J. Liu, W. Wu, H. Zhou, and B. Zhuang, "Ptqd: Accurate post-training quantization for diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[167] T. Salimans and J. Ho, "Progressive Distillation for Fast Sampling of Diffusion Models," Jun. 2022, arXiv:2202.00512 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2202.00512

[168] J. Li, W. Feng, T.-J. Fu, X. Wang, S. Basu, W. Chen, and W. Y. Wang, "T2V-Turbo: Breaking the Quality Bottleneck of Video Consistency Model with Mixed Reward Feedback," May 2024, arXiv:2405.18750 [cs]. [Online]. Available: http://arxiv.org/abs/2405.18750

[169] G. Team, "Mochi," 2024.

[170] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[171] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[172] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[173] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, "Pick-a-pic: An open dataset of user preferences for text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 36 652–36 663, 2023.

[174] X. He, D. Jiang, G. Zhang, M. Ku, A. Soni, S. Siu, H. Chen, A. Chandra, Z. Jiang, A. Arulraj, K. Wang, Q. D. Do, Y. Ni, B. Lyu, Y. Narsupalli, R. Fan, Z. Lyu, Y. Lin, and W. Chen, "VideoScore: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation," Jun. 2024, arXiv:2406.15252 [cs]. [Online]. Available: http://arxiv.org/abs/2406.15252

[175] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, 2021.

[176] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5148–5157. [Online]. Available: http://openaccess.thecvf.com/content/ICCV2021/html/Ke_MUSIQ_Multi-Scale_Image_Quality_Transformer_ICCV_2021_paper.html

[177] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, and P. Luo, "Mvbench: A comprehensive multi-modal video understanding benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 195–22 206. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Li_MVBench_A_Comprehensive_Multi-modal_Video_Understanding_Benchmark_CVPR_2024_paper.html

[178] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 144–20 154.

[179] S. Yuan, J. Huang, Y. Xu, Y. Liu, S. Zhang, Y. Shi, R. Zhu, X. Cheng, J. Luo, and L. Yuan, "ChronoMagic-Bench: A Benchmark for Metamorphic Evaluation of Text-to-Time-lapse Video Generation," Jun. 2024, arXiv:2406.18522 [cs]. [Online]. Available: http://arxiv.org/abs/2406.18522

[180] W. Ren, H. Yang, G. Zhang, C. Wei, X. Du, W. Huang, and W. Chen, "ConsistI2V: Enhancing Visual Consistency for Image-to-Video Generation," Jun. 2024, arXiv:2402.04324 [cs]. [Online]. Available: http://arxiv.org/abs/2402.04324

[181] Z. Zhang, X. Li, W. Sun, J. Jia, X. Min, Z. Zhang, C. Li, Z. Chen, P. Wang, Z. Ji, F. Sun, S. Jui, and G. Zhai, "Benchmarking AIGC Video Quality Assessment: A Dataset and Unified Model," Jul. 2024, arXiv:2407.21408 [cs]. [Online]. Available: http://arxiv.org/abs/2407.21408

[182] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, and N. Chanpaisit, "Vbench: Comprehensive benchmark suite for video generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 807–21 818. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html

[183] Y. Liu, X. Cun, X. Liu, X. Wang, Y. Zhang, H. Chen, Y. Liu, T. Zeng, R. Chan, and Y. Shan, "Evalcrafter: Benchmarking and evaluating large video generation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 139–22 149.

[184] M. Liao, H. Lu, X. Zhang, F. Wan, T. Wang, Y. Zhao, W. Zuo, Q. Ye, and J. Wang, "Evaluation of Text-to-Video Generation Models: A Dynamics Perspective," Jul. 2024, arXiv:2407.01094 [cs]. [Online]. Available: http://arxiv.org/abs/2407.01094

[185] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu *et al.*, "Emu3: Next-token prediction is all you need," *arXiv preprint arXiv:2409.18869*, 2024.

[186] M. Farré, A. Marafioti, L. Tunstall, L. Von Werra, and T. Wolf, "Finevideo," https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024.

[187] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2630–2640. [Online]. Available: http://openaccess.thecvf.com/content_ICCV_2019/html/Miech_HowTo100M_Learning_a_Text-Video_Embedding_by_Watching_Hundred_Million_Narrated_ICCV_2019_paper.html

[188] H. Xue, T. Hang, Y. Zeng, Y. Sun, B. Liu, H. Yang, J. Fu, and B. Guo, "Advancing high-resolution video-language representation with large-scale video transcriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5036–5045. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2022/html/Xue_Advancing_High-Resolution_Video-Language_Representation_With_Large-Scale_Video_Transcriptions_CVPR_2022_paper.html

[189] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1728–1738. [Online]. Available: http://openaccess.thecvf.com/content/ICCV2021/html/Bain_Frozen_in_Time_A_Joint_Video_and_Image_Encoder_for_ICCV_2021_paper.html

[190] K. Soomro, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[191] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *arXiv preprint arXiv:1808.01340*, 2018.

[192] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[193] W. Xiong, W. Luo, L. Ma, W. Liu, and J. Luo, "Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2364–2373.

[194] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[195] H. Sim, J. Oh, and M. Kim, "Xvfi: extreme video frame interpolation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14 489–14 498.

[196] T.-S. Chen, A. Siarohin, W. Menapace, E. Deyneka, H.-w. Chao, B. E. Jeon, Y. Fang, H.-Y. Lee, J. Ren, M.-H. Yang *et al.*, "Panda-70m: Captioning 70m videos with multiple cross-modality teachers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 320–13 331.

[197] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *International Journal of Computer Vision*, vol. 123, pp. 94–120, 2017.

[198] H. Xu, Q. Ye, X. Wu, M. Yan, Y. Miao, J. Ye, G. Xu, A. Hu, Y. Shi, G. Xu *et al.*, "Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks," *arXiv preprint arXiv:2306.04362*, 2023.

[199] Z. Tan, X. Yang, L. Qin, and H. Li, "Vidgen-1m: A large-scale dataset for text-to-video generation," *arXiv preprint arXiv:2408.02629*, 2024.

[200] W. Wang and Y. Yang, "VidProM: A Million-scale Real Prompt-Gallery Dataset for Text-to-Video Diffusion Models," May 2024, arXiv:2403.06098 [cs]. [Online]. Available: http://arxiv.org/abs/2403.06098

[201] H. Chen, Y. Hong, Z. Huang, Z. Xu, Z. Gu, Y. Li, J. Lan, H. Zhu, J. Zhang, W. Wang, and H. Li, "DeMamba: AI-Generated Video Detection on Million-Scale GenVideo Benchmark," Jul. 2024, arXiv:2405.19707 [cs]. [Online]. Available: http://arxiv.org/abs/2405.19707

[202] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft COCO Captions: Data Collection and Evaluation Server," Apr. 2015, arXiv:1504.00325 [cs]. [Online]. Available: http://arxiv.org/abs/1504.00325

[203] M. Dehghani, B. Mustafa, J. Djolonga, J. Heek, M. Minderer, M. Caron, A. Steiner, J. Puigcerver, R. Geirhos, and I. M. Alabdulmohsin, "Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution," *Advances in Neural Information Processing Systems*, vol. 36, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/06ea400b9b7cfce6428ec27a371632eb-Abstract-Conference.html

[204] Z. Lu, Z. Wang, D. Huang, C. Wu, X. Liu, W. Ouyang, and L. Bai, "FiT: Flexible Vision Transformer for Diffusion Model," Feb. 2024, arXiv:2402.12376 [cs]. [Online]. Available: http://arxiv.org/abs/2402.12376

[205] J. Lv, Y. Huang, M. Yan, J. Huang, J. Liu, Y. Liu, Y. Wen, X. Chen, and S. Chen, "GPT4Motion: Scripting Physical Motions in Text-to-Video Generation via Blender-Oriented GPT Planning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1430–1440. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024W/PBDL/html/Lv_GPT4Motion_Scripting_Physical_Motions_in_Text-to-Video_Generation_via_Blender-Oriented_GPT_CVPRW_2024_paper.html

[206] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, "The stable signature: Rooting watermarks in latent diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 466–22 477. [Online]. Available: http://openaccess.thecvf.com/content/ICCV2023/html/Fernandez_The_Stable_Signature_Rooting_Watermarks_in_Latent_Diffusion_Models_ICCV_2023_paper.html

[207] S. Dathathri, A. See, S. Ghaisas, P.-S. Huang, R. McAdam, J. Welbl, V. Bachani, A. Kaskasoli, R. Stanforth, and T. Matejovicova, "Scalable watermarking for identifying large language model outputs," *Nature*, vol. 634, no. 8035, pp. 818–823, 2024, publisher: Nature Publishing Group UK London. [Online]. Available: https://www.nature.com/articles/s41586-024-08025-4

TABLE IX
COMPARISON OF STATE-OF-THE-ART MODELS ON VBENCH ON CHALLENGING METRICS

Model Name	Semantic Score	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class	Multiple Objects	Human Action	Color	Spatial Relationship	Scene
T2V-Turbo-v2	77.12	90.00	62.61	71.78	95.33	61.49	96.2	92.53	43.32	56.40
MiniMax-Video-01	77.65	64.91	63.03	67.17	87.83	76.04	92.4	90.36	75.50	50.68
Gen-3	75.17	60.14	63.34	66.82	87.81	53.64	96.4	80.90	65.09	54.57
Vchitect-2.0 (VEnhancer)	77.06	63.89	60.41	65.35	86.61	68.84	97.2	87.04	57.55	56.57
Kling	75.68	46.94	61.21	65.62	87.24	68.05	93.4	89.90	73.03	50.86
LaVie-2	75.76	31.11	67.62	70.39	97.52	64.88	96.4	91.65	38.68	49.59
Pyramid Flow	69.62	64.63	63.26	65.01	86.67	50.71	85.60	82.87	59.53	43.20
CogVideoX-5B	77.04	70.97	61.98	62.90	85.23	62.11	99.40	82.81	66.35	53.20
Allegro	72.98	55.00	63.74	63.60	87.52	59.92	91.40	82.77	67.15	46.72
Emu3	68.43	79.27	59.64	62.63	86.17	44.64	77.71	88.34	68.73	37.11
CogVideoX-2B	75.83	59.86	60.82	61.68	83.37	62.63	98.00	79.41	69.90	51.14
Pika-1.0	71.77	47.50	62.04	61.87	88.72	43.08	86.20	90.57	61.03	49.83
Gen-2	73.03	18.89	66.96	67.42	90.92	55.47	89.20	89.49	66.91	48.91
VideoCrafter-2.0	73.42	42.50	63.13	67.22	92.55	40.66	95.00	92.92	35.86	55.29
AnimateDiff-V2	69.75	40.83	67.16	70.10	90.90	36.88	92.60	87.47	34.60	50.19
OpenSora V1.2	73.39	42.39	56.85	63.34	82.22	51.83	91.20	90.08	68.56	42.44
Variance	8.55	326.73	7.99	9.70	16.55	124.26	31.39	18.77	185.76	28.45

Elnaz Soleimani is a Machine Learning researcher at Vikit.ai, specializing in computer vision and generative artificial intelligence. She obtained her Master's degree in Artificial Intelligence from Tehran Polytechnic in 2018. Since then, she has been actively involved in research and development within the field of computer vision. Her current research interests include advancing generative AI techniques and their applications in video and image generation.

Dr. Ghazaleh Khodabandelou is an Associate Professor at the University of Paris-Est, with a Ph.D. from Paris Sorbonne University. She has collaborated with esteemed institutions, including Paderborn University, DePaul University, and the University of Washington. Her research focuses on using Advanced Optimization Techniques, Large Language Models, and Vision-Language Models to create AI-driven solutions in Human-Centric Computing, intention mining, Ambient Assisted Living, healthcare, robotics, and emotional intelligence.

APPENDIX
BENCHMARKING

Table IX summarizes the performance of state-of-the-art models on VBench's challenging quality metrics, emphasizing areas where model behaviors diverge most. For example, the dynamic degree metric, which assesses the motion level in generated videos, reveals a substantial performance gap among models. It appears that models often prioritize consistency over motion, as even completely static videos can achieve high scores on other temporal consistency metrics.

(a)

(b)

(c)

(d)

(e)

Fig. 9. Example of failed cases in state-of-the-art video generation models. Video a, b-top, c, and d are generated by Runway. Video b-bottom is generated by CogVideoX-2B and video e is generated by Haiper.