

A Survey on Video Diffusion Models

ZHEN XING, QIJUN FENG, and HAORAN CHEN, Shanghai Collaborative Innovation Center of Intelligent Visual Computing and School of Computer Science, Fudan University, China

QI DAI and HAN HU, Microsoft Research Asia, China

HANG XU, Noah's Ark Lab, China

ZUXUAN WU* and YU-GANG JIANG*, Shanghai Collaborative Innovation Center of Intelligent Visual Computing and School of Computer Science, Fudan University, China

The recent wave of AI-generated content (AIGC) has witnessed substantial success in computer vision, with the diffusion model playing a crucial role in this achievement. Due to their impressive generative capabilities, diffusion models are gradually superseding methods based on GANs and auto-regressive Transformers, demonstrating exceptional performance not only in image generation and editing, but also in the realm of video-related research. However, existing surveys mainly focus on diffusion models in the context of image generation, with few up-to-date reviews on their application in the video domain. To address this gap, this paper presents a comprehensive review of video diffusion models in the AIGC era. Specifically, we begin with a concise introduction to the fundamentals and evolution of diffusion models. Subsequently, we present an overview of research on diffusion models in the video domain, categorizing the work into three key areas: video generation, video editing, and other video understanding tasks. We conduct a thorough review of the literature in these three key areas, including further categorization and practical contributions in the field. Finally, we discuss the challenges faced by research in this domain and outline potential future developmental trends. A comprehensive list of video diffusion models studied in this survey is available at <https://github.com/ChenHsing/Awesome-Video-Diffusion-Models>.

CCS Concepts: • Computing methodologies → Computer vision.

Additional Key Words and Phrases: Survey, Video Diffusion Model, Video Generation, Video Editing, AIGC

ACM Reference Format:

Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. 2024. A Survey on Video Diffusion Models. *J. ACM* 37, 4, Article 1 (August 2024), 37 pages. <https://doi.org/XXXXXXX-XXXXXX>

1 Introduction

AI-generated content (AIGC) is currently one of the most prominent research fields in computer vision and artificial intelligence. It has not only garnered extensive attention and scholarly investigation, but also exerted profound influence across industries and other applications, such as computer graphics, art and design, medical imaging, *etc.* Among these endeavors, a series of approaches represented by diffusion models [83, 161, 188, 193, 197, 198, 297] have emerged as particularly successful,

*Corresponding author

Authors' Contact Information: Zhen Xing, zxing20@fudan.edu.cn; Qijun Feng, qjfeng21@m.fudan.edu.cn; Haoran Chen, chenhran21@m.fudan.edu.cn, Shanghai Collaborative Innovation Center of Intelligent Visual Computing and School of Computer Science, Fudan University, China; Qi Dai, qid@microsoft.com; Han Hu, ancientmooner@gmail.com, Microsoft Research Asia, China; Hang Xu, Noah's Ark Lab, China; Zuxuan Wu, zxwu@fudan.edu.cn; Yu-Gang Jiang, ygi@fudan.edu.cn, Shanghai Collaborative Innovation Center of Intelligent Visual Computing and School of Computer Science, Fudan University, China.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 1557-735X/2024/8-ART1

<https://doi.org/XXXXXXX.XXXXXXX>

rapidly supplanting methods based on generative adversarial networks (GANs) [67, 109–111, 203] and auto-regressive Transformers [41, 49, 189, 285] to become the predominant approach for image generation. Due to their strong controllability, photorealistic generation, and impressive diversity, diffusion-based methods also bloom across a wide range of computer vision tasks, including image editing [14, 80, 160, 226], dense prediction [22, 74, 102, 168, 236], and diverse areas such as video synthesis [81, 85, 154, 207, 244, 266] and 3D generation [54, 137, 153, 182]. As one of the most important mediums, video emerges as a dominant force on the Internet. Compared to mere text and static image, video presents a trove of dynamic information, providing users with a more comprehensive and immersive visual experience. Research on video tasks based on the diffusion models is progressively gaining traction. As shown in Fig. 1, the number of research publications of video diffusion models has increased significantly since 2022 and can be categorized into three major classes: video generation [11, 85, 113, 154, 207, 244, 266], video editing [48, 56, 136, 183, 262], and video understanding [25, 130, 166, 237].

With the rapid advancement of video diffusion models [85] and their demonstration of impressive results, the endeavor to track and compare recent research on this topic gains great importance. Several survey articles have covered foundational models in the era of AIGC [260, 294], encompassing the diffusion model itself [229, 276] and multi-modal learning [59, 292, 309]. There are also surveys specifically focusing on text-to-image [293] research and text-to-3D [128] applications. However, these surveys either provide only a coarse coverage of the video diffusion models or place greater emphasis on image models [229, 292, 293]. As such, in this work, we aim to fulfill the blank with a comprehensive review of the methodologies, experimental settings, benchmark datasets, and other video applications of the diffusion model.

Contribution: In this survey, we systematically track and summarize recent literature concerning video diffusion models, encompassing domains such as video generation, editing, and other aspects of video understanding. By extracting shared technical details, this survey covers the most representative works in the field. Background and relevant knowledge preliminaries concerning video diffusion models are also introduced. Furthermore, we conduct a comprehensive analysis and comparison of benchmarks and settings for video generation. To the best of our knowledge, we are the first to concentrate on this specific domain. More importantly, given the rapid evolution of the video diffusion, we might not cover all the latest advancements in this survey. Therefore we encourage researchers to get in touch with us to share their new findings in this domain, enabling us to maintain currency. These novel contributions will be incorporated into the revised version for discussion.

Survey Pipeline: In Section 2, we will cover background knowledge, including problem definition, datasets, evaluation metrics, and relevant research domains. Subsequently, in Section 3, we primarily present an overview of methods in the field of video generation. In Section 4, we delve into the principal studies concerning video editing tasks. In Section 5, we elucidate the various directions of utilizing diffusion models for video understanding. In Section 6, we highlight the existing research challenges and potential future avenues, culminating in our concluding remarks in Section 7.

2 Preliminaries

In this section, we first present preliminaries of diffusion models, followed by reviewing the related research domains. Finally, we introduce the commonly used datasets and evaluation metrics.

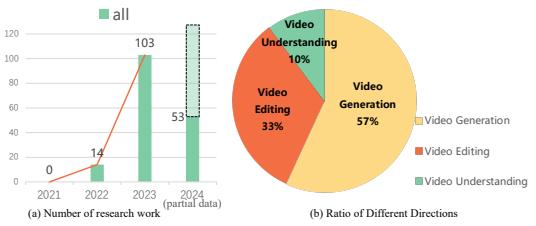


Fig. 1. Summarization on video diffusion model research works. (a) The number of related research works is rapidly increasing. (b) Video generation and editing are the top two research areas using diffusion models.

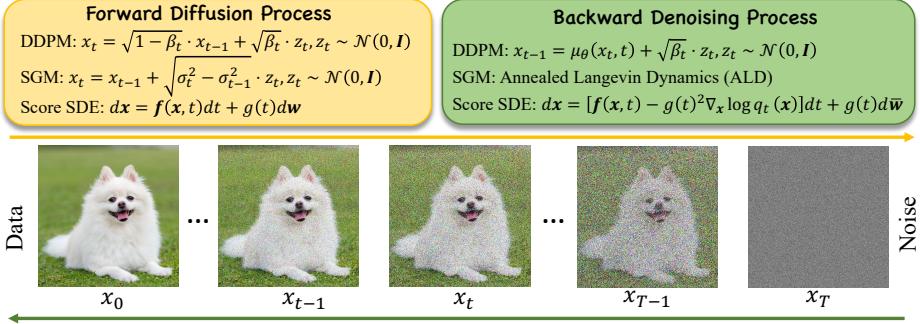


Fig. 2. **Overview of diffusion model.** We demonstrate the diffusion and denoising process of DDPM, SGM and Score SDE.

2.1 Diffusion Model

Diffusion models [82, 213] are a category of probabilistic generative models that learn to reverse a process that gradually degrades the training data structure and have become the new state-of-the-art family of deep generative models. They have broken the long-held dominance of generative adversarial networks (GANs) [66] in a variety of challenging tasks such as image generation [171, 194, 209, 211, 212, 232], image super-resolution [8, 112, 194, 199], and image editing [4, 34]. Current research on diffusion models is mostly based on three predominant formulations: denoising diffusion probabilistic models (DDPMs) [82, 172, 209], score-based generative models (SGMs) [211, 212], and stochastic differential equations (Score SDEs) [210, 213]. The diffusion and denoising processes of these three formulations are summarized and demonstrated in Fig. 2.

2.1.1 Denoising Diffusion Probabilistic Models (DDPMs). A *denoising diffusion probabilistic model* (DDPM) [82, 172, 209] involves two Markov chains: a forward chain that perturbs data to noise, and a reverse chain that converts noise back to data. The former aims at transforming any data into a simple prior distribution, while the latter learns transition kernels to reverse the former process. New data points can be generated by first sampling a random vector from the prior distribution, followed by ancestral sampling through the reverse Markov chain. The pivot of this sampling process is to train the reverse Markov chain to match the actual time reversal of the forward Markov chain.

Formally, given a data distribution $x_0 \sim q(x_0)$, the forward Markov process generates a sequence of random variables x_1, x_2, \dots, x_T with transition kernel $q(x_t|x_{t-1})$. The joint distribution of x_1, x_2, \dots, x_T conditioned on x_0 , denoted as $q(x_1, \dots, x_T|x_0)$, can be factorized into

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

Typically, the transition kernel is designed as

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t \mathbf{I}), \quad (2)$$

where $\beta_t \in (0, 1)$ is a hyperparameter chosen ahead of model training.

The reverse Markov chain is parameterized by a prior distribution $p(x_T) = \mathcal{N}(x_T; 0, \mathbf{I})$ and a learnable transition kernel $p_\theta(x_{t-1}|x_t)$ which takes the form of

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

where θ denotes model parameters and the mean $\mu_\theta(x_t, t)$ and variance $\Sigma_\theta(x_t, t)$ are parameterized by deep neural networks. With the reverse Markov chain, we can generate new data x_0 by first

sampling a noise vector $x_T \sim p(x_T)$, then iteratively sampling from the learnable transition kernel $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$ until $t = 1$.

2.1.2 Score-Based Generative Models (SGMs). The key idea of score-based generative models (SGMs) [211, 212] is to perturb data using various levels of noise and simultaneously estimate the scores corresponding to all noise levels by training a single conditional score network. Samples are generated by chaining the score functions at decreasing noise levels with score-based sampling approaches. Training and sampling are entirely decoupled in the formulation of SGMs.

With similar notations in Sec. 2.1.1, let $q(x_0)$ be the data distribution, and $0 < \sigma_1 < \sigma_2 < \dots < \sigma_T$ be a sequence of noise levels. A typical example of SGMs involves perturbing a data point x_0 to x_t by the Gaussian noise distribution $q(x_t|x_0) = \mathcal{N}(x_t; x_0, \sigma_t^2 I)$, which yields a sequence of noisy data densities $q(x_1), q(x_2), \dots, q(x_T)$, where $q(x_t) := \int q(x_t)q(x_0)dx_0$. A noise-conditional score network (NCSN) is a deep neural network $s_\theta(x, t)$ trained to estimate the score function $\nabla_{x_t} \log q(x_t)$. We can directly employ techniques such as score matching, denoising score matching, and sliced score matching to train our NCSN from perturbed data points.

For sample generation, SGMs leverage iterative approaches to produce samples from $s_\theta(x, T), s_\theta(x, T-1), \dots, s_\theta(x, 0)$ in succession by using techniques such as annealed Langevin dynamics (ALD).

2.1.3 Stochastic Differential Equations (Score SDEs). Perturbing data with multiple noise scales is key to the success of the above methods. Score SDEs [213] generalize this idea further to an infinite number of noise scales. The diffusion process can be modeled as the solution to the following stochastic differential equation (SDE):

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w} \quad (4)$$

where $\mathbf{f}(\mathbf{x}, t)$ and $g(t)$ are diffusion and drift functions of the SDE, and \mathbf{w} is a standard Wiener process.

Starting from samples of $\mathbf{x}(T) \sim p_T$ and reversing the process, we can obtain samples $\mathbf{x}(0) \sim p_0$ through this reverse-time SDE:

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log q_t(\mathbf{x})]dt + g(t)d\bar{\mathbf{w}} \quad (5)$$

where $\bar{\mathbf{w}}$ is a standard Wiener process when time flows backwards. Once the score of each marginal distribution, $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$, is known for all t , we can derive the reverse diffusion process from Eq.(5) and simulate it to sample from p_0 .

2.2 Related Tasks

The applications of video diffusion model contain a wide scope of video tasks, including video generation, video editing, and various other forms of video understanding. The methodologies for these tasks share similarities, often formulating the problems as diffusion generation tasks or utilizing the potent controlled generation capabilities of diffusion models for downstream tasks. In this survey, the main focus lies on the tasks such as Text-to-Video generation [81, 207, 266], unconditional video generation [91, 152, 157], and text-guided video editing [48, 165, 262], etc.

- **Text-to-Video Generation** aims to automatically generate corresponding videos based on the textual descriptions. This typically involves comprehending the scenes, objects, and actions within the textual descriptions and translating them into a sequence of coherent visual frames, resulting in a video with both logical and visual consistency. T2V has broad applications, including the automatic generation of movies [311], animations [75, 78], virtual reality content, educational demonstration videos [277], etc.

- **Unconditional Video Generation** is a generative task where the objective is to generate a continuous and visually coherent sequence of videos starting from random noise or a fixed initial state,

Dataset	Year	Text	Domain	#Clips	Resolution
MSR-VTT [271]	2016	Manual	Open	10K	240P
DideMo [3]	2017	Manual	Flickr	27K	-
LSMDC [192]	2017	Manual	Movie	118K	1080P
ActivityNet [119]	2017	Manual	Action	100K	-
YouCook2 [307]	2018	Manual	Cooking	14K	-
How2 [202]	2018	Manual	Instruct	80K	-
VATEX [245]	2019	Manual	Action	41K	240P
HowTo100M [162]	2019	ASR	Instruct	136M	240P
WTS70M [217]	2020	Metadata	Action	70M	-
YT-Temporal [290]	2021	ASR	Open	180M	-
WebVid10M [5]	2021	Alt-text	Open	10.7M	360P
Echo-Dynamic [191]	2021	Manual	Echocardiogram	10K	-
Tiktok [241]	2021	Manual	Action	0.3K	-
HD-VILA [273]	2022	ASR	Open	103M	720P
VideoCC3M [167]	2022	Transfer	Open	10.3M	-
HD-VG-130M [244]	2023	Generated	Open	130M	720P
InternVid [250]	2023	Generated	Open	234M	720P
CelebV-Text [286]	2023	Generated	Face	70K	480P
Panda-70M [28]	2024	Generated	Open	70.8M	720P

Table 1. The comparison of main caption-level video datasets.

without relying on specific input conditions. Unlike conditional video generation, unconditional video generation does not require any external guidance or prior information [79, 85, 154]. The generative model needs to learn how to capture temporal dynamics, actions, and visual coherence in the absence of explicit inputs, to produce video content that is both realistic and diverse. This is crucial for exploring the ability of generative models to learn video content from unsupervised data and showcase diversity.

- **Text-guided Video Editing** is a technique that involves using textual descriptions to guide the process of editing video content. In this task, a natural language description is provided as input, describing the desired changes or modifications to be applied to a video. The system then analyzes the textual input, extracts relevant information such as objects, actions, or scenes, and uses this information to guide the editing process. Text-guided video editing offers a way to facilitate efficient and intuitive editing by allowing editors to communicate their intentions using natural language [48, 154, 205], potentially reducing the need for manual and time-consuming frame-by-frame editing.

2.3 Datasets and Metrics

2.3.1 Data.

The evolution of video understanding tasks often aligns with the development of video datasets, and the same applies to video generation tasks. In the early stages of video generation, tasks are limited to training on low-resolution [215], small-scale datasets to specific domains [206, 269], resulting in relatively monotonous video generation. With the emergence of large-scale video-text paired datasets, tasks such as general text-to-video generation [85, 207] began to gain traction. Thus, the datasets of video generation can be mainly categorized into caption-level and category-level, as will be discussed separately.

- **Caption-level Datasets** consist of videos paired with descriptive text captions, providing essential data for training models to generate videos based on textual descriptions. We list several common caption-level datasets in Table 1, which vary in scale and domain. Early caption-level video datasets were primarily used for video-text retrieval tasks [3, 192, 271], with small-scales (less than 120K) and a limited focus on specific domains (*e.g.* movie [192], action [119, 217], cooking [307]). With the introduction of the open-domain WebVid-10M [5] dataset, a new task of text-to-video (T2V) generation gains momentum, leading researchers to focus on open-domain T2V generation tasks. Despite being a mainstream benchmark dataset for T2V tasks, it still suffers from issues such as low resolution (360P) and watermarked content. Subsequently, to enhance the resolution and broader

Datasets	Year	Categories	#Clips	Resolution
UCF-101 [214]	2012	101	13K	256 × 256
Cityscapes [37]	2015	30	3K	256 × 256
Moving MNIST [215]	2016	10	10K	64 × 64
Kinetics-400 [18]	2017	400	260K	256 × 256
BAIR [45]	2017	2	45K	64 × 64
DAVIS [181]	2017	-	90	1280 × 720
Sky Time-Lapse [269]	2018	1	38K	256 × 256
Sstv2 [68]	2018	174	220K	256 × 256
Kinetics-600 [17]	2018	600	495K	256 × 256
Epic-Kitchen [39]	2018	149	90K	1920 × 1080
Tai-Chi-HD [206]	2019	1	3K	256 × 256
Bridge Data [46]	2021	10	7K	256 × 256
Mountain Bike [13]	2022	1	1K	576 × 1024
RDS [11]	2023	2	683K	512 × 1024

Table 2. The comparison of existing category-level datasets for video generation and editing.

coverage of videos in the general text-to-video (T2V) tasks, Panda-70M [28], VideoFactory [244] and InternVid [250] introduce larger-scale (70M & 130M & 234M) and high-definition (720P) open-domain datasets. To collect diverse video datasets, VidRD [71] utilizes static images [204], long videos [298] and short videos [18] when constructing the video-text dataset.

- **Category-level Datasets** consist of videos grouped into specific categories, with each video labeled by its category. The datasets are commonly utilized for unconditional video generation or class conditional video generation tasks. We summarize category-level commonly used video datasets in Table 2. Notably, several of these datasets are also applied to other tasks. For instance, UCF-101 [214], Kinetics [17, 18], and Something-Something [68] are typical benchmarks for action recognition. DAVIS [181] was initially proposed for the video object segmentation task and later became a commonly used benchmark for video editing. Among these datasets, UCF-101 [214] stands out as the most widely utilized in video generation, serving as a benchmark for unconditional video generation, category-based conditional generation, and video prediction applications. It comprises samples from YouTube [283] that encompasses 101 action categories, including human sports, musical instrument playing, and interactive actions. Akin to UCF, Kinetics-400 [18] and Kinetics-600 [17] are two datasets encompassing more complex action categories and larger data scale, while retaining the same application scope as UCF-101 [214]. The Something-Something [68] dataset, on the other hand, possesses both category-level and caption-level labels, rendering it particularly suitable for text-conditional video prediction tasks [73]. It is noteworthy that these sizable datasets that originally played pivotal roles in the realm of action recognition exhibit smaller scales (less than 50K) and single-category [206, 269], single-domain attributes (digital number [215], driving scenery [11, 13, 37], egocentric [39], robot [46]) and is thereby inadequate for producing high-quality videos. Consequently, in recent years, datasets specifically crafted for video generation tasks are proposed, typically originating from featuring unique attributes, such as high resolution (1080P) [11] or extended duration [13, 282]. For example, Long Video GAN [13] proposes dataset which has 66 videos with an average duration of 6504 frames at 30fps. Video LDM [11] collects RDS dataset consisting of 683,060 real driving videos of 8 seconds in length each with 1080P resolution.

2.3.2 Evaluation Metrics.

Evaluation metrics for video generation are commonly categorized into quantitative and qualitative measures. For qualitative measures, human subjective evaluation has been used in several works [207, 244, 262, 266], where evaluators are typically presented with two or more generated videos to compare against videos synthesized by other competitive models. Observers generally engage in voting-based assessments regarding the realism, natural coherence, and text alignment of the videos (T2V tasks). However, human evaluation is both costly and at the risk of failing to reflect the full capabilities of the model [1]. Therefore, in the following we will primarily delve into the quantitative evaluation standards for image-level and video-level assessments.

- **Image-level Metrics** Videos are composed of a sequence of image frames, thus image-level evaluation metrics can provide a certain amount of insight into the quality of the generated video frames. Commonly employed image-level metrics include Fréchet Inception Distance (FID) [179], Peak Signal-to-Noise Ratio (PSNR) [255], Structural Similarity Index (SSIM) [255], and CLIPSIM [186]. FID [179] assesses the quality of generated videos by comparing synthesized video frames to real video frames. It involves preprocessing the images for normalization to a consistent scale, utilizing InceptionV3 [219] to extract features from real and synthesized videos, and computing mean and covariance matrices. These statistics are then combined to calculate the FID [179] score.

Both SSIM [255] and PSNR [255] are pixel-level metrics. SSIM [255] evaluates brightness, contrast, and structural features of original and generated images, while PSNR [255] is a coefficient representing the ratio between peak signal and Mean Squared Error (MSE) [254]. These two metrics are commonly used to assess the quality of reconstructed image frames, and are applied in tasks such as super-resolution and in-painting. CLIPSIM [186] is a method for measuring image-text relevance. Based on the CLIP [186] model, it extracts both image and text features and then computes the similarity between them. This metric is often employed in text-conditional video generation or editing tasks [11, 106, 207, 244, 262, 266].

- **Video-level Metrics** Although image-level evaluation metrics represent the quality of generated video frames, they primarily focus on individual frames, disregarding the temporal coherence of the video. Video-level metrics, on the other hand, would provide a more comprehensive evaluation of video generation. Fréchet Video Distance (FVD) [231] is a video quality evaluation metric based on FID [179]. Unlike image-level methods that use the Inception [219] network to extract features from a single frame, FVD [231] employs the Inflated-3D Convnets (I3D) [18] pre-trained on Kinetics [18] to extract features from video clips. Subsequently, FVD scores are computed through the combination of means and covariance matrices. Kernel Video Distance (KVD) [230] is also based on I3D [18] features, but it differentiates itself by utilizing Maximum Mean Discrepancy (MMD) [70], a kernel-based method, to assess the quality of generated videos. Video IS (Inception Score) [200] calculates the Inception score of generated videos using features extracted by the 3D-Convnets (C3D) [223], which is often applied in evaluation on UCF-101 [214]. High-quality videos are characterized by a low entropy probability, denoted as $P(y|x)$, whereas diversity is assessed by examining the marginal distribution across all videos, which should exhibit a high level of entropy. Frame Consistency CLIP Score [186] is often used in video editing tasks [262, 266, 299] to measure the coherence of edited videos. It is calculated by obtaining CLIP image embeddings for all frames and averaging the cosine similarity between all pairs of frames.

3 Video Generation

In this section, we categorize video generation into four groups and provide detailed reviews for each: Text-to-video (T2V) generation (Sec. 3.1), Video Generation with other conditions (Sec. 3.2), Unconditional Video Generation (Sec. 3.3) and Video Completion (Sec. 3.4). Finally, we summarize the settings and evaluation metrics, and present a comprehensive comparison of various models in Sec. 3.5. The taxonomy details of video generation are demonstrated in Fig. 3.

3.1 Video Generation with Text Condition

Evidenced by recent research [174, 188, 193], the interaction between generative AI and natural language is of paramount importance. While significant progress has been achieved in generating images from text [41, 161, 188, 193], the development of Text-to-Video (T2V) approaches is still in its early stages. In this section, we will introduce training-based and training-free T2V methods, respectively.

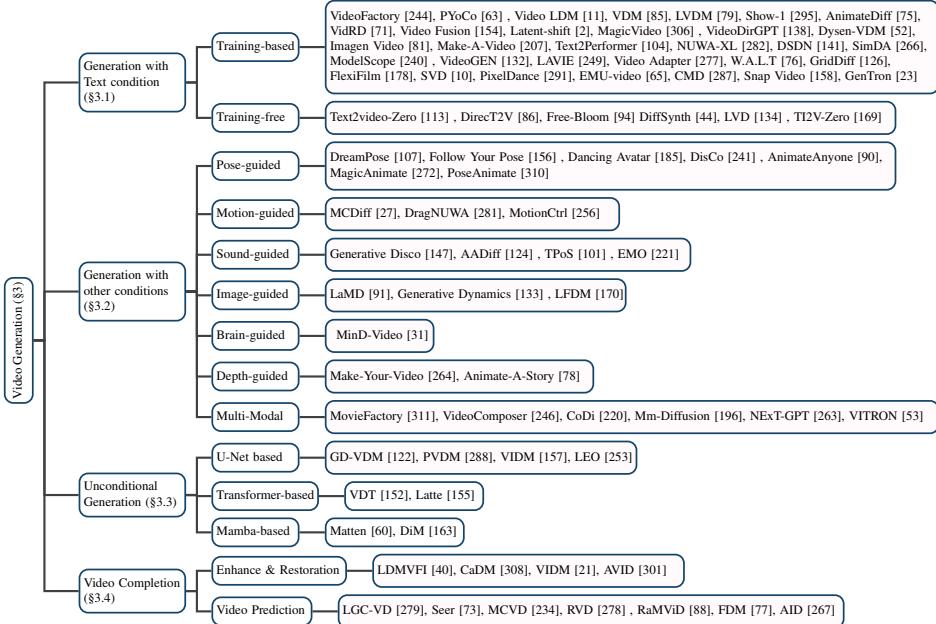


Fig. 3. Taxonomy of Video Generation. Key aspects of Video Generation include General T2V Generation, Domain-specific Generation, Conditional Control Generation, and Video Completion.

3.1.1 Training-based T2V Diffusion Methods. In the preceding discussion, we have briefly recapitulated a few T2V methods that do not rely on the diffusion model. Moving forward, we predominantly introduce the utilization of the currently most prominent diffusion model in the realm of T2V tasks.

• **Early T2V Exploration** Among the multitude of endeavors, VDM [85] stands as the pioneer in devising a video diffusion model for video generation. It extends the conventional image diffusion U-Net [195] architecture to a 3D U-Net structure and employs joint training with both images and videos. The conditional sampling technique it employs enables generating videos of enhanced quality and extended duration. Being the first exploration of a diffusion model for T2V, it also accommodates tasks such as unconditional generation and video prediction.

In contrast to VDM [85], which requires paired *video-text* datasets, Make-A-Video [207] introduces a novel paradigm. Here, the network learns visual-textual correlations from paired *image-text* data and captures video motion from unsupervised video data. This innovative approach reduces the reliance on data collection, resulting in the generation of diverse and realistic videos. Furthermore, by employing multiple super-resolution models and interpolation networks, it achieves higher-definition and frame-rate generated videos.

• **Temporal Modeling Exploration** While previous approaches leverage diffusion at pixel-level, MagicVideo [306] stands as one of the earliest works to employ the Latent Diffusion Model (LDM) [193] for T2V generation in latent space. By utilizing diffusion models in a lower-dimensional latent space, it significantly reduces computational complexity, thereby accelerating processing speed. The introduced frame-wise lightweight adaptor aligns the distributions of images and videos so that the proposed directed attention can better model temporal relationships to ensure video consistency.

Concurrently, LVDM [79] also employs the LDM [193] as its backbone, utilizing a hierarchical framework to model the latent space. By employing a mask sampling technique, the model becomes capable of generating longer videos. It incorporates techniques such as Conditional Latent Perturbation [83] and Unconditional Guidance [84] to mitigate performance degradation in the later stages of

auto-regressive generation tasks. With this training approach, it can be applied to video prediction tasks, even generating long videos consisting of thousands of frames.

VideoFactory [244] introduces a swapped cross-attention mechanism to facilitate interaction between the temporal and spatial modules, resulting in improved temporal relationship modeling. Besides, trained on its proposed HD-VG-130M dataset, the approach presented in the paper is capable of generating high-resolution videos at (1376×768) resolution.

ModelScope [240] incorporates spatial-temporal convolution and attention into LDM [193] for T2V tasks. It adopts a mixed training approach using LAION [204] and WebVid [5], and serves as an open-source baseline method.

Previous methods predominantly rely on 1D convolutions or temporal attention [306] to establish temporal relationships. Latent-Shift [2], on the other hand, focuses on lightweight temporal modeling. Drawing inspiration from TSM [139], it shifts channels between adjacent frames in convolution blocks for temporal modeling. Additionally, the model maintains the original T2I [193] capability while generating videos.

- **Multi-stage T2V methods** Imagen Video [81] extends the T2I model, Imagen [198], for video generation using a cascaded video diffusion model composed of seven sub-models: one for base video generation, three for spatial super-resolution, and three for temporal super-resolution. This three-stage training pipeline employs T2I techniques like classifier-free guidance [84], conditioning augmentation [83], and v-parameterization [201]. Progressive distillation techniques [159, 201] are used to speed up sampling time, making these multi-stage training techniques effective for high-definition video generation.

Concurrently, Video LDM [11] trains a T2V network composed of three training stages, including key-frame T2V generation, video frame interpolation and spatial super-resolution modules. It adds temporal attention layer and 3D convolution layer to the spatial layer, enabling the generation of key frames in the first stage. Subsequently, through the implementation of a mask sampling method, a frame interpolation model is trained, extending key frames of short videos to higher frame rates. Lastly, a video super-resolution model is employed to enhance the resolution.

Similarly, LAVIE [249] employs a cascaded video diffusion model composed of three stages: a base T2V stage, a temporal interpolation stage, and a video super-resolution stage. Furthermore, it validates that the process of joint image-video fine-tuning can yield high-quality and creative outcomes.

Show-1 [295] introduces the fusion of pixel-based [121] and latent-based [193] diffusion models for T2V generation. Its framework has four stages: key frame generation, frame interpolation, and super-resolution at a low-resolution pixel level, followed by a latent super-resolution module to enhance video resolution cost-effectively. Pixel-level stages ensure precise text alignment in the videos. Latent-level stage offers a cost-effective means of enhancing video resolution.

- **Noise Prior Exploration** While most of the methods mentioned denoising each frame independently through diffusion models, VideoFusion [154] stands out by considering the content redundancy and temporal correlations among different frames. Specifically, it decomposes the diffusion process using a shared base noise for each frame and residual noise along the temporal axis. This noise decomposition is achieved through two co-training networks. Such an approach is introduced to ensure consistency in generating frame motion, although it may lead to limited diversity. Furthermore, the paper shows that employing T2I backbones [188] for training T2V models accelerates convergence, but its text embedding might face challenges in understanding long temporal sequences of text.

PYoCo [63] acknowledges that directly extending the image noise prior to video can yield suboptimal outcomes in T2V tasks. As a solution, it intricately devises a video noise prior and fine-tune the eDiff-I [6] model for video generation. The proposed noise prior involves sampling correlated noise

for different frames within the video. The authors validate that the proposed mixed and progressive noise models are better suited for T2V tasks.

FlexiFilm [178] introduces a temporal conditioner and a resampling strategy to maintain consistency between multimodal conditions and the generated videos. These strategies enhance consistency between generated videos and multimodal conditions, and address overexposure to produce videos up to 30 seconds long.

VidRD [71] introduces the Reuse and Diffuse framework, which iteratively generates additional frames by reusing the original latent representations and following the previous diffusion process. In this way, it can generate long videos.

- **Efficient Training** ED-T2V [145] utilizes LDM [193] as its backbone and freezes a substantial portion of parameters to reduce training costs. It introduces identity attention and temporal cross-attention to ensure temporal coherence. The approach proposed in this paper manages to lower training costs while maintaining comparable T2V generation performance.

SimDA [266] devises a parameter-efficient training approach for T2V tasks by maintaining the parameter of T2I model [193] fixed. It incorporates a lightweight spatial adapter for transferring visual information for T2V learning. Additionally, it introduces a temporal adapter to model temporal relationships in lower feature dimensions. The proposed latent shift attention aids in maintaining video consistency. Moreover, the lightweight architecture enables speed up inference and makes it adaptable for video editing tasks.

To reduce the substantial computational resource requirements of 3D U-Nets, GridDiff [126] proposes a method for video generation using a 2D U-Net. Specifically, this approach treats video as a grid image. This method allows for the straightforward extension of image generation and editing techniques to the video domain.

- **Personalized Video Generation** Personalized video generation generally refers to creating videos tailored to a specific protagonist or style, addressing the generation of videos customized for personal preferences or characteristics. AnimateDiff [75] notices the success of LoRA [89] and Dreambooth [197] in personalized T2I models and aims to extend their effectiveness to video animation. Furthermore, the authors aim at training a model that can be adapted to generate diverse personalized videos, without the need of repeatedly retraining on video datasets. This involves using a T2I model as a base generator and adding a motion module to learn motion dynamics. During inference, the personalized T2I model can replace the base T2I weights, enabling personalized video generation.

- **Image-conditioned T2V methods** To address the issue of flickers and artifacts in T2V-generated videos, DSDN [141] introduces a dual-stream diffusion model, one for video content and the other for motion. In this way, it can maintain a strong alignment between content and motion. By decomposing the video generation process into content and motion components, it is possible to generate continuous videos with fewer flickers.

VideoGen [132] first utilizes a T2I model [193] to generate images based on the text prompt, which serves as a reference image for guiding video generation. Subsequently, an efficient cascaded latent diffusion module is introduced, employing flow-based temporal upsampling steps to enhance temporal resolution.

Additionally, MicroCinema [248], PixelDance [291], and EMU-VIDEO [65] all follow the image-conditioned T2V pipeline. They inject image control conditions into the T2V generation process through an additional Appearance-conditioned network or conditioned latent concatenation. Utilizing a reference image improves visual fidelity and reduces artifacts, allowing the model to focus more on learning video dynamics.

SVD [10] validates the data scaling capability of the video diffusion model. It collects and labels hundreds of millions of video-text data, and releases the image2video video generation model, which has been utilized in many subsequent works.

- **Complex Dynamics Modeling** The generation of Text-to-Video (T2V) encounters challenges in modeling complex dynamics, particularly regarding disruptions in action coherence. To address this, Dysen-VDM [52] introduces a method that transforms textual information into dynamic scene graphs. Leveraging Large Language Model (LLM) [174], Dysen-VDM [52] identifies pivotal actions from the input text and arranges them chronologically, enriching scenes with pertinent descriptive details. Furthermore, the model benefits from in-context learning of LLM, endowing it with robust spatio-temporal modeling. This approach demonstrates remarkable superiority in the synthesis of complex actions.

VideoDirGPT [138] also utilizes LLM to plan the generation of video content. For a given text input, it is expanded into a video plan through GPT-4 [175], which includes scene descriptions, entities along with their layouts, and the distribution of entities within backgrounds. Subsequently, corresponding videos are generated by the model with explicit control over layouts. This approach demonstrates layout and motion control advantages for complex dynamic video generation.

- **Domain-specific T2V Generation** Video-Adapter [277] introduces a novel setting by transferring pre-trained general T2V models to domain-specific T2V tasks. By decomposing the domain-specific video distribution into pretrained noise and a small training component, it substantially reduces the cost of transferring training. The efficacy of this approach is verified in T2V generation for Ego4D [69] and Bridge Data [46] scenarios.

NUWA-XL [282] employs a coarse-to-fine generative paradigm, facilitating parallel video generation. It initially employs global diffusion to generate key frames, followed by utilizing a local diffusion model to interpolate between two frames. This methodology enables the creation of lengthy videos spanning up to 3376 frames, thus establishing a benchmark for the generation of animations. This work focuses on the field of cartoon video generation, utilizing its techniques to produce cartoon videos lasting several minutes.

Text2Performer [104] decomposes human-centric videos into appearance and motion representations. It first employs unsupervised training on natural human videos using a VQVAE [233] latent space to disentangle appearance and pose representations. Subsequently, it utilizes a continuous VQ-diffuser [12, 72] to sample continuous pose embeddings. Finally, it employs a motion-aware masking strategy in the spatio-temporal domain on the pose embeddings to enhance temporal correlations.

- **Transformer-based T2V Generation** The emergence of Sora [176] significantly boosts the popularity of video diffusion models, utilizing the Diffusion Transformer architecture to achieve substantial improvements in video stability and consistency. W.A.L.T [76] explores Transformer-based video diffusion models, using a causal encoder to compress images and videos into a unified space for joint training of the diffusion model. Additionally, it adopts a window attention architecture [151] tailored for joint spatial and spatio-temporal generative modeling, while the employed cascading super-resolution network can generate stable videos at high resolutions.

CMD [287] proposes a video diffusion model that separates content and motion, first training an autoencoder to encode videos into content frames and motion latents. Then, it uses the Diffusion Transformer(DiT) [180] architecture to generate motion vectors. This method enables the generation of high-resolution videos at a low computational cost.

Snap Video [158] proposes a modification to the EDM [108] diffusion framework for generating high-resolution videos and treats images as high frame-rate videos to avoid image-video modality mismatches. It replaces the U-Net architecture with a transformer architecture [24], scaling up to billions of parameters to demonstrate competitive results.

GenTron [23] adapts Diffusion Transformers (DiTs) [180] from class to text conditioning, a process that involves thorough empirical exploration of the conditioning mechanism. Additionally, it proposes a motion-free guidance method to modulate the weight of motion information in the generated video.

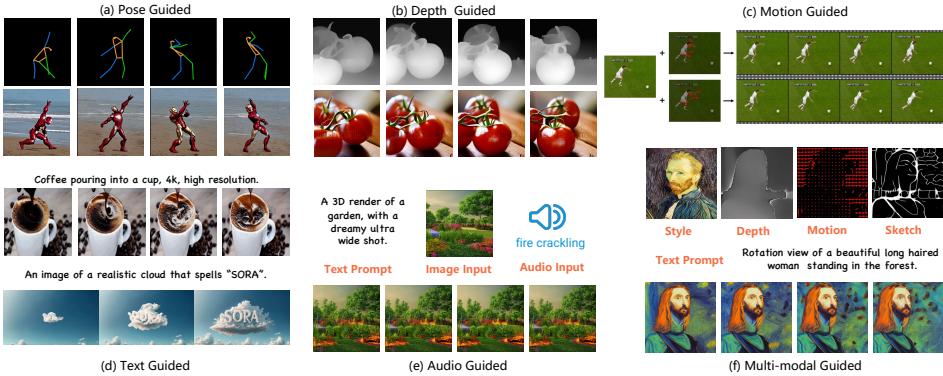


Fig. 4. Conditional video generation results with (a) Pose Guided [156], (b) Depth Guided [264], (c) Motion Guided [27], (d) Text Guided [176, 266], (e) Audio Guided [101] and (f) Multi-modal Guided [246].

3.1.2 Training-free T2V Diffusion Methods. While former methods are all training-based T2V approaches that typically rely on extensive datasets like WebVid [5] or other video datasets [250, 273]. Some recent researches [94, 113] aim at reducing heavy training costs by developing training-free T2V approaches, as will be introduced next.

Text2Video-Zero [113] utilizes the pre-trained T2I model Stable Diffusion [193] for video synthesis. To maintain consistency across different frames, it performs a Cross-Attention mechanism between each frame and the first frame. Additionally, it enriches motion dynamics by modifying the sampling method of latent code. Moreover, this method can be combined with conditional generation and editing techniques such as ControlNet [297] and Instruct-Pix2Pix [14], enabling the controlled generation of videos.

DirecT2V [86] and Free-Bloom [94], on the other hand, introduce large language model (LLM) [175, 177] to generate frame-to-frame descriptions based on a single abstract user prompt. LLM directors are employed to breakdown user input into frame-level descriptions. Additionally, to maintain continuity between frames, DirecT2V [86] uses a novel value mapping and dual-softmax filtering approach. Free-Bloom [94] proposes a series of reverse process enhancements, which encompass joint noise sampling, step-aware attention shifting, and dual-path interpolation. Experimental results demonstrate these modifications enhance the zero-shot video generation capabilities.

To handle intricate spatial-temporal prompts, LVD [134] first utilizes LLM [175] to generate dynamic scene layouts and then employs these layouts to guide video generation. Its approach requires no training and guides video diffusion models by adjusting attention maps based on the layouts, enabling the generation of complex dynamic videos.

DiffSynth [44] introduces a latent in-iteration deflickering framework and a video deflickering algorithm to reduce flickering and produce coherent videos. It is applicable to various domains, including video stylization and 3D rendering.

T2V-Zero [169] proposes a tuning-free method that incorporates image-conditioned control constraints to the T2V method, enabling more controllable video generation. It employs a "repeat-and-slide" strategy along with DDPM [82] inversion to inject the initial frame as a control condition into the video generation process, which is an effective approach for long video generation.

3.2 Video Generation with other Conditions

Most of the previously introduced methods pertain to text-to-video generation. In this subsection, we focus on video generation conditioned on other modalities (*e.g.* pose, sound and depth). We show the condition-controlled video generation examples in Fig. 4.

3.2.1 Pose-guided Video Generation. Follow Your Pose [156] presents a video generation model driven by pose and text control. It employs a two-stage training process by utilizing image-pose pairs and pose-free videos. In the first stage, a T2I model is finetuned using (image, pose) pairs, enabling pose-controlled generation. In the second stage, the model leverages unlabeled videos to learn temporal modeling by incorporating temporal modules. This two-stage training imparts the model with both pose control and temporal modeling capabilities.

Dreampose [107] constructs a dual-path CLIP-VAE [186] image encoder and adapter module to replace the original CLIP text encoder in LDM [193] as the conditioning component. Given a single human image and a pose sequence, this study can generate a corresponding human pose video based on the provided pose information.

Dancing Avatar [185] focuses on synthesizing human dance videos. It utilizes a T2I model [193] to generate each frame of the video in an auto-regressive manner. To ensure consistency throughout the entire video, a frame alignment module combined with insights from ChatGPT [174] is utilized to enhance coherence between adjacent frames. Additionally, it leverages OpenPose ControlNet [297] to harness the ability to generate high-quality human body videos based on poses.

Disco [241] addresses the novel problem of referring human dance generation using ControlNet [297] for background control, Grounded-SAM [118] for foreground extraction, and OpenPose [15] for pose skeleton extraction. Large-scale image datasets [98, 140, 204] are used for human attribute pre-training, creating a strong foundation for human-specific video generation tasks.

Animate Anyone [90] and MagicAnimate [272] use character images as reference frames and design a pose encoder to inject sequence pose information into the image-to-video generation, enabling characters to follow predefined poses for dancing. They demonstrate significant zero-shot capabilities with only a small training set, and the industrial applications prove the tremendous potential of video diffusion models.

PoseAnimate [310] proposes a zero-shot I2V framework for character animation. It introduces a Pose-Aware Control Module to control the pose, a Dual Consistency Attention Module to ensure temporal consistency, and a Mask-Guided Decoupling Module (MGDM) to refine distinct feature perception. Through these modules, the model demonstrates zero-shot pose animation capabilities.

3.2.2 Motion-guided Video Generation. MCDiff [27] is the pioneer in considering motion as a condition for controlling video synthesis. The approach involves providing the first frame of a video along with a sequence of stroke motions. Initially, a flow completion model [239] is utilized to predict dense video motion based on sparse stroke motion control. Subsequently, the model employs an auto-regressive approach using the dense motion map to predict subsequent frames, ultimately resulting in the synthesis of a complete video.

DragNUWA [281] simultaneously introduces text, image, and trajectory information to provide fine-grained control over video content from semantic, spatial and temporal perspectives. To further address the lack of open-domain trajectory control in previous works, the authors proposed a Trajectory Sampler (TS) to enable open-domain control of arbitrary trajectories, a Multiscale Fusion (MF) to control trajectories in different granularities, and an Adaptive Training (AT) strategy to generate consistent video following trajectories.

MotionCtrl [256] proposes two control modules, one for camera motion and the other for object motion. It suggests methods for dataset collection tailored to these types of motion, and the trained control modules can be effectively integrated into various base models of video diffusion to enhance motion control.

3.2.3 Sound-guided Video Generation. AADiff [124] introduces the concept of using audio and text together as conditions for video synthesis. The approach starts by separately encoding text and audio using dedicated encoders [47]. Then, the similarity between the text and audio embeddings

is computed, and the text token with the highest similarity is selected. This selected text token is used in a prompt2prompt [80] fashion to edit frames. This approach enables the generation of audio-synchronized videos without requiring any additional training.

Generative Disco [147] is designed for text-to-video generation aimed at music visualization. The system employs a pipeline that involves a large language model [175] followed by a text-to-image model [193] to achieve its goals.

TPoS [101] integrates audio inputs with variable temporal semantics and magnitude, building upon the LDM [193] to extend the utilization of audio modality in generative models. This approach outperforms widely-used audio-to-video benchmarks, as demonstrated by objective evaluations and user studies, highlighting its superior performance.

EMO [221] proposes an audio-to-video framework for human talking generation, extracting audio features through an audio encoder and then injecting them into the diffusion model using audio attention. This allows for the creation of highly expressive and lifelike animations without the need for 3D facial information.

3.2.4 Image-guided Video Generation. LaMD [91] first trains an autoencoder to separate motion information within videos. Then a diffusion-based motion generator is trained to generate video motion. Through this methodology, guided by motion, the model achieves the capability to generate high-quality perceptual videos given the first frame.

LFDM [170] leverages conditional images and text for human-centric video generation. Firstly a latent flow auto-encoder is trained to reconstruct videos. Moreover, a flow predictor [242] can be employed in intermediary steps to predict flow motion. Subsequently, in the second stage, a diffusion model is trained with image, flow, and text prompts as conditions to generate coherent videos.

Generative Dynamics [133] presents an approach to modeling scene dynamics in image space. It extracts motion trajectories from real video sequences exhibiting natural motion. For a single image, the diffusion model, through a frequency-coordinated diffusion sampling process, predicts a long-term motion representation in the Fourier domain for each pixel. This representation can be converted into dense motion trajectories spanning the entire video. When combined with an image rendering module, it enables the transformation of static images into seamless looping dynamic videos, facilitating realistic user interactions with the depicted objects.

3.2.5 Brain-guided Video Generation. MinD-Video [31] is the pioneering effort to explore video generation through continuous fMRI data. The approach begins by aligning MRI data with images and text using contrastive learning. Next, a trained MRI encoder replaces the CLIP text encoder as the input for conditioning. This is further enhanced through the design of a temporal attention module to model sequence dynamics. The resultant model is capable of reconstructing videos that possess precise semantics, motions, and scene dynamics, surpassing groundtruth performance and setting a new benchmark in this field.

3.2.6 Depth-guided Video Generation. Make-Your-Video [264] employs a novel approach for text-depth condition video generation. It integrates depth information as a conditioning factor by extracting it using MiDAS [190] during training. Additionally, the method introduces a causal attention mask to facilitate the synthesis of longer videos. Comparisons with state-of-the-art techniques demonstrate the method's superiority in controllable text-to-video generation, showcasing better quantitative and qualitative performance.

In Animate-A-Story [78], an innovative approach is introduced that divides video generation into two steps. The first step, Motion Structure Retrieval, involves retrieving the most relevant videos from a large video database based on a given text prompt [5]. Depth maps of these retrieved videos are obtained using offline depth estimation methods [190], which then serve as motion guidance. In

the second step, Structure-Guided Text-to-Video Synthesis is employed to train a video generation model guided by the structural motion derived from the depth maps. Such a two-step approach enables the creation of personalized videos based on customized text descriptions.

3.2.7 Multi-modal guided Video Generation. VideoComposer [246] focuses on video generation conditioned on multi-modal, encompassing textual, spatial, and temporal conditions. Specifically, it introduces a Spatio-Temporal Condition encoder that allows flexible combinations of various conditions. This ultimately enables the incorporation of multiple modalities, such as sketch, mask, depth, and motion vectors. By harnessing control from multiple modalities, VideoComposer [246] achieves higher video quality and improved detail in the generated content.

MM-Diffusion [196] represents the inaugural endeavor in a joint audio-video generation. To realize the generation of multimodal content, it introduces a bifurcated architecture comprising two subnets tasked with video and audio generation, respectively. To ensure coherence between the outputs of these two subnets, a random-shift based attention block has been devised to establish interconnections. Beyond its capacity for unconditional audio-video generation, MM-Diffusion [196] also exhibits pronounced aptitude in effectuating video-to-audio translation.

MovieFactory [311] is dedicated to applying the diffusion model to the generation of film-style videos. It leverages ChatGPT [174, 187] to elaborate on user-provided text, creating comprehensive sequential scripts for movie generation. In addition, an audio retrieval system has been devised to provide voice overs for videos. Through the aforementioned techniques, the realization of generating multi-modal audio-visual content is achieved.

CoDi [220] presents a novel generative model that possesses the capability of creating diverse combinations of output modalities, encompassing language, images, videos, or audio, from varying combinations of input modalities. This is achieved by constructing a shared multimodal space, facilitating the generation of arbitrary modality combinations through the alignment of input and output spaces across diverse modalities.

NExT-GPT [263] presents an end-to-end, any-to-any multimodal LLM system. It integrates LLM [32] with multimodal adapters and diverse diffusion decoders, enabling the system to perceive input in arbitrary combinations of text, images, videos, and audio, and generate corresponding output.

VITRON [53], utilizing an LLM backbone [32], incorporates various encoders for images and videos, enhancing the versatility of the model and supporting several tasks within a unified framework. Particularly, it achieves good results in both image-to-video and text-to-video settings, demonstrating the tremendous potential of combining LLMs with video diffusion models.

3.3 Unconditional Video Generation

In this section, we delve into unconditional video generation. It refers to generating videos that belong to the specific domain without extra conditions. The focal points of these studies revolve around the design of video representations and the architecture of diffusion model networks.

- **U-Net based Generation** As one of the earliest works on unconditional video diffusion models and a significant baseline method, VIDM [157] uses two streams: the content generation stream for video frame content generation and the motion stream to define video motion. By merging these streams, consistent videos are generated. Additionally, the authors employ Positional Group Normalization (PosGN) [172] to enhance video continuity and explore combining Implicit Motion Condition (IMC) and PosGN to address the generation consistency of long videos.

Similar to LDM [193], PVDM [288] first trains an auto-encoder to map pixels into a lower-dimensional latent space, followed by applying a diffusion denoising generative model in the latent space to synthesize videos. This approach reduces both training and inference costs while capable of maintaining satisfactory generation quality.

GD-VDM [122] first generates depth map videos where scene and layout generation are prioritized whereas fine details and textures are abstracted away. Then, the generated depth maps are provided as a conditioning signal to further generate the remaining details of the video. This methodology retains superior detail generation capabilities and is particularly applicable to complex driving scene video generation tasks.

LEO [253] involves representing motion within the generation process through a sequence of flow maps, thereby inherently separating motion from appearance. It achieves human video generation through the combination of a flow-based image animator and a Latent Motion Diffusion Model. The former learns the reconstruction from flow maps to motion codes, while the latter captures motion priors to obtain motion codes. The synergy of these two methods enables effective learning of human video correlations. Furthermore, this approach can be extended to tasks such as infinite-length human video synthesis and content-preserving video editing.

- **Transformer-based Generation** Inspired by the success of DiT [180] in image generation, VDT [152] and Latte [155] aim to extend the diffusion transformer to the video generation tasks. They explore several spatio-temporal attention variants and conditional injection modules, while also validating the scaling capabilities of the video diffusion transformer.
- **Mamba-based Generation** Matten [60] proposes a video diffusion model architecture that combines Mamba and Attention, exploring several spatio-temporal modeling variants based on Mamba and Attention, validating that Mamba can be applied to video diffusion models. DiM [163] explores a pure Mamba structure for image and video diffusion models, validating its scalability and ability to generate high-quality videos at a lower computational cost.

3.4 Video Completion

Video completion constitutes a pivotal task within the realm of video generation. In the subsequent sections, we will delineate the distinct facets of video enhancement and restoration and video prediction.

3.4.1 Video Enhancement and Restoration. CaDM [308] introduces a novel Neural-enhanced Video Streaming paradigm aimed at substantially diminishing streaming delivery bitrates, all the while maintaining a notably heightened restoration capability in contrast to prevailing methodologies. Primarily, the proposed CaDM [308] approach improves the compression efficacy of the encoder through the concurrent reduction of frame resolution and color bit-depth in video streams. Furthermore, CaDM [308] empowers the decoder with superior enhancement capabilities by imbuing the denoising diffusion restoration process with an awareness of the resolution-color conditions stipulated by the encoder.

LDMVFI [40] is the first to use a conditional latent diffusion model for video frame interpolation (VFI). This approach introduces several innovative concepts, including a VFI-specific autoencoding network with efficient self-attention modules and deformable kernel-based frame synthesis techniques to improve performance.

VIDM [21] utilizes the pre-trained LDM [193] for video inpainting. By providing a mask for first-person perspective videos, it leverages the image completion capabilities of LDM to generate inpainted videos.

AVID [301] first achieves fixed-length video in-painting by designing a motion module and structure guidance. Then, through the designed Temporal MultiDiffusion sampling pipeline with a middle-frame attention guidance mechanism, it extends in-painting to videos of arbitrary length.

3.4.2 Video Prediction. Seer [73] is dedicated to the exploration of the text-guided video prediction task. It leverages the Latent Diffusion Model (LDM) as its foundational backbone. Through the

integration of spatial-temporal attention within an auto-regressive framework, alongside the implementation of the Frame Sequential Text Decomposer module, Seer adeptly transfers the knowledge priors of Text-to-Image (T2I) models to the domain of video prediction. This migration has led to substantial performance enhancements, notably demonstrated on benchmarks [46, 68].

FDM [77] introduces a novel hierarchy sampling scheme for the purpose of long video prediction tasks. Additionally, a new CARLA [42] dataset is proposed. In comparison to auto-regressive methods, the proposed approach is not only more efficient but also yields superior generative outcomes.

MCVD [234] employs a probabilistic conditional score-based denoising diffusion model for both unconditional generation and interpolation tasks. The introduced masking approach is capable of masking all past or future frames, thereby enabling the prediction of frames from either the past or the future. Additionally, it adopts an autoregressive approach to generate videos of variable lengths in a block-wise fashion.

LGC-VD [279] introduces a Local-Global Context guided Video Diffusion model designed to encompass diverse perceptual conditions. LGC-VD employs a two-stage training approach and treats prediction errors as a form of data augmentation. This strategy effectively addresses prediction errors and notably reinforces stability in the context of long video prediction tasks.

RVD [278] adopts a diffusion model that utilizes the context vector of a convolutional Recurrent Neural Network as conditions to generate a residual, which is then added to a deterministic next-frame prediction. The authors demonstrate that employing residual prediction is more effective than directly predicting future frames.

RaMViD [88] employs 3D convolutions to extend the image diffusion model into the realm of video tasks. It introduces a novel conditional training technique and utilizes a mask condition to extend its applicability to various completion tasks, including video prediction [17, 45], infilling [45], and upsampling [17, 214].

AID [267] first proposed transferring a general video diffusion model to specific domain video prediction tasks. It introduces the DQFormer architecture, which injects the initial frame, user instructions, and multi-modal large model [143] planning predictions into the video generation process. Additionally, it utilizes a lightweight adapter for efficient transfer SVD [10] to robotics and first-person video generation.

3.5 Benchmark Results

This section systematically compares various video generation methods under zero-shot and finetuned settings. For each setting, we start by introducing their commonly used datasets. Subsequently, we state the detailed evaluation metrics utilized for each of the datasets. Finally, we present a comprehensive performances comparison of the methods.

3.5.1 Zero-shot T2V Generation. • **Datasets.** General T2V methods, such as Make-A-Video [207] and VideoLDM [11], are primarily evaluated on the MSRVTT [271] and UCF-101 [214] datasets in a zero-shot manner. MSRVTT [271] is a video retrieval dataset, where each video clip is accompanied by approximately 20 natural sentences for description. Typically, the textual descriptions corresponding to the 2,990 video clips in its test set are utilized as prompts to produce the corresponding generated videos. UCF-101 [214] is an action recognition dataset with 101 action categories. In the context of T2V models, videos are typically generated based on the category names or manually set prompts corresponding to these action categories.

• **Evaluation Metrics.** When evaluating under the zero-shot setting, it is common practice to assess video quality using FVD [231] and FID [179] metrics on the MSRVTT [271] dataset. CLIPSIM [186] is used to measure the alignment between text and video. For the UCF-101 [214] dataset, the typical

Method	Year	Training Data	Extra Dependency	Resolution	Params(B)	MSRVTT [271]			UCF-101 [214]		
						FID(↓)	FVD(↓)	CLIPSIM(↑)	FID(↓)	FVD(↓)	IS(↑)
Non-diffusion based method											
CogVideo [87]	2022	[5](5.4M)	-	256 × 256	15.5	23.59	1294	0.2631	179.00	701.59	25.27
MMVG [58]	2023	[5](2.5M)	-	256 × 256	-	-	-	0.2644	-	-	-
Diffusion based method											
LVDM [79]	2022	[5](2M)	-	256 × 256	1.16	-	742	0.2381	-	641.8	-
MagicVideo [306]	2022	[5](10M)	-	256 × 256	-	-	998	-	145.00	699.00	-
Make-A-Video [207]	2022	[5, 273]	-	256 × 256	9.72	13.17	-	0.3049	-	367.23	33.00
ED-T2V [145]	2023	[5](10M)	-	256 × 256	1.30	-	-	0.2763	-	-	-
InternVid [250]	2023	[5](10M)+18M*	-	256 × 256	-	-	-	0.2951	60.25	616.51	21.04
Video-LDM [11]	2023	[5](10M)	-	256 × 256	4.20	-	-	0.2929	-	550.61	33.45
VideoComposer [246]	2023	[5](10M)	-	256 × 256	1.85	-	580	0.2932	-	-	-
Latent-shift [2]	2023	[5](10M)	-	256 × 256	1.53	15.23	-	0.2773	-	-	-
VideoFusion [154]	2023	[5](10M)	-	256 × 256	1.83	-	581	0.2795	75.77	639.90	17.49
Make-Your-Video [264]	2023	[5](10M)	Depth Input	256 × 256	-	-	-	-	-	330.49	-
PYOCo [63]	2023	[5](22.5M)		256 × 256	-	9.73	-	-	-	355.19	47.76
CoDi [220]	2023	[5, 273]	-	512 × 512	-	-	-	0.2890	-	-	-
NEXT-GPT [263]	2023	[5, 273]	-	320 × 576	1.83	13.04	-	0.3085	-	-	-
SimDA [266]	2023	[5](10M)	ChatGPT	256 × 256	1.08	-	456	0.2945	-	-	-
Dysen-VDM [52]	2023	[5](10M)		256 × 256	-	12.64	-	0.3204	-	325.42	35.57
VideoFactory [244]	2023	[5, 273]	-	256 × 256	2.04	-	-	0.3005	-	410.00	-
ModelScope [240]	2023	[5](10M)	-	256 × 256	1.70	11.09	550	0.2930	-	410.00	-
VideoGen [132]	2023	[5](10M)	Reference Image	256 × 256	-	-	-	0.3127	-	554.00	71.61
Animate-A-Story [78]	2023	[5](10M)		256 × 256	-	-	-	-	-	515.15	-
VidRID [71]	2023	[5, 18, 298](5.3M*)	-	256 × 256	-	-	-	-	-	363.19	39.37
LAVIE [249]	2023	[5](10M)+25M*	-	320 × 512	3.00	-	-	0.2949	-	526.30	-
VideoDirGPT [138]	2023	[5](10M)	GPT-4	256 × 256	1.92	12.22	550	0.2860	-	-	-
Show-1 [295]	2023	[5](10M)		320 × 576	-	13.08	538	0.3072	-	394.46	35.42
Dynamicrafter [265]	2023	[5](10M)	Reference Image	256 × 256	-	-	234	-	-	429.23	-
EMU-Video [65]	2023	34M*		256 × 256	-	-	-	-	-	606.20	42.70
PixelDance [291]	2023	[5](10M)+50W*	Reference Image	256 × 256	1.50	-	381	0.3125	49.36	242.82	42.10
MicroCinema [248]	2023	[5](10M)		256 × 256	2.42	-	377	0.2967	-	342.86	37.46
ART-V [248]	2023	[5](5M)	Reference Image	256 × 256	-	-	291	0.2859	-	315.69	50.34
SVD [10]	2023	577M*		256 × 384	-	-	-	-	-	242.02	-
W.A.L.T [76]	2023	89M*	-	128 × 128	3.00	-	244.7	-	-	258.1	35.1
CMD [287]	2023	[5](10M)	Reference Image	512 × 1024	1.60	-	-	0.2894	-	504.00	-
Video [158]	2024	238k hours		288 × 512	3.90	9.35	104.0	0.2793	28.1	200.20	38.89

Table 3. Zero-shot Text-to-Video generation comparison on MSRVTT [271] and UCF-101 [214] dataset. We report the Fréchet Video Distance (FVD) scores, CLIPSIM scores, Fréchet Image Distance (FID) and Inception Score (IS). The dataset marked with “*” indicates the use of a self-collected dataset.

evaluation metrics include Inception Score [200], FVD [231], and FID [179] to evaluate the quality of generated videos and their frames.

• **Results Comparison.** In Table 3, we present the zero-shot performance of current general T2V methods on MSRVTT [271] and UCF-101 [214]. We also provide information about their parameter number, training data, extra dependencies, and resolution. It can be observed that methods relying on ChatGPT [52] or other input conditions [78, 264] exhibit a significant advantage over others, and the utilization of additional data [207, 244, 250] often leads to improved performance.

3.5.2 Finetuned Video Generation. • **Datasets.** Finetuned video generation methods refer to generating videos after fine-tuning on a specific dataset. This typically includes unconditional video generation and class conditional video generation. It primarily focuses on three specific datasets: UCF-101 [214], Taichi-HD [206], and Time-lapse [269]. These datasets are associated with distinct domains: UCF-101 concentrates on human sports, Taichi-HD mainly comprises Tai Chi videos, and Time-lapse predominantly features time-lapse footage of the sky. Additionally, there are several other benchmarks available [18, 37, 45], but we choose these three as they are the most commonly used ones.

• **Evaluation Metrics.** In the evaluation of the Finetuned Video Generation task, commonly used metrics for the UCF-101 [214] dataset include IS [200] (Inception Score) and FVD [231] (Fréchet Video Distance). For the Time-lapse [269] and Taichi-HD [206] datasets, common evaluation metrics include FVD and KVD [230].

• **Results Comparison.** In Table 4, we present the performance of current state-of-the-art methods fine-tuned on benchmark datasets. Similarly, further details regarding the method type, resolution,

Method	Year	Type	Resolution	Extra	UCF-101 [214]		Taichi-HD [206]		Time-lapse [269]	
					FVD(↓)	IS(↑)	FVD(↓)	KVD(↓)	FVD(↓)	KVD(↓)
MoCoGAN [225]	2018	GAN	64 × 64	-	-	12.42	-	-	206.6	-
TGANv2 [200]	2020	GAN	128 × 128	-	-	26.60	-	-	-	-
StyleGAN-V [208]	2022	GAN	256 × 256	-	-	23.94	-	-	79.52	-
MoCoGAN-HD [222]	2021	GAN	256 × 256	-	700	33.95	144.7	25.4	183.6	13.9
DIGAN [289]	2022	GAN	128 × 128	-	577	32.70	128.1	20.6	114.6	6.8
StyleInv [251]	2023	GAN	256 × 256	-	-	-	186.72	-	77.04	-
MMVG [58]	2023	VQGAN	128 × 128	-	-	58.3	395	-	-	-
VideoGPT [275]	2021	Autoregressive	64 × 64	-	-	24.69	-	-	222.7	-
CCV [123]	2021	Autoregressive	128 × 128	-	386	24.47	-	-	-	-
TATS [61]	2022	Autoregressive	128 × 128	Class Condition	278	79.28	94.6	9.8	132.6	5.7
CogVideo [87]	2022	Autoregressive	160 × 160	Pretrain+Class Condition	626	50.46	-	-	-	-
VDM [85]	2022	Diffusion	64 × 64	-	-	57.80	-	-	-	-
LVDM [79]	2022	Diffusion	256 × 256	-	372	27.00	99	15.3	95.2	3.9
VIDM [157]	2022	Diffusion	256 × 256	-	294.7	-	121.9	-	57.4	-
LEO [253]	2022	Diffusion	256 × 256	-	-	-	122.7	20.49	57.4	-
VideoFusion [154]	2023	Diffusion	128 × 128	-	220	72.22	56.4	6.9	47.0	5.3
PVDM [288]	2023	Diffusion	256 × 256	-	343.6	74.40	-	-	55.41	-
VDT [152]	2023	Diffusion	64 × 64	-	283.0	-	-	-	-	-
PYeoCo [63]	2023	Diffusion	256 × 256	-	310	60.01	-	-	-	-
Dysen-VDM [52]	2023	Diffusion	256 × 256	ChatGPT	255.42	95.23	-	-	-	-
Latent-Shift [2]	2022	Diffusion	256 × 256	Class Condition	360	92.72	-	-	-	-
ED-T2V [145]	2023	Diffusion	256 × 256	Class Condition	320	83.36	-	-	-	-
Make-A-Video [207]	2023	Diffusion	256 × 256	Pretrain+Class Condition	81.25	82.55	-	-	-	-
VideoGen [132]	2023	Diffusion	256 × 256	Pretrain+Class Condition	345	82.78	-	-	-	-
Matten [60]	2024	Diffusion	256 × 256	-	210.61	-	158.56	-	53.56	-
Dim [163]	2024	Diffusion	256 × 256	-	206.83	-	-	-	-	-
Latte [155]	2024	Diffusion	256 × 256	-	333.61	73.31	97.09	-	42.67	-
AID [267]	2024	Diffusion	256 × 256	Pretrain+Class Condition	102	-	-	-	-	-

Table 4. Finetuned video generated results of UCF-101 [214], Taichi-HD [206] and Time-lapse [269]. We report the FVD, IS and KVD scores evaluation metric of clips with 16 frames. We also report the resolution of each video frame for each evaluation result.

and extra dependencies are provided. It is evident that diffusion-based methods exhibit a significant advantage compared to traditional GANs [222, 225, 289] and autoregressive Transformer [87, 275] methods. Furthermore, if there is a large-scale pretraining or class conditioning, the performance tends to be further enhanced.

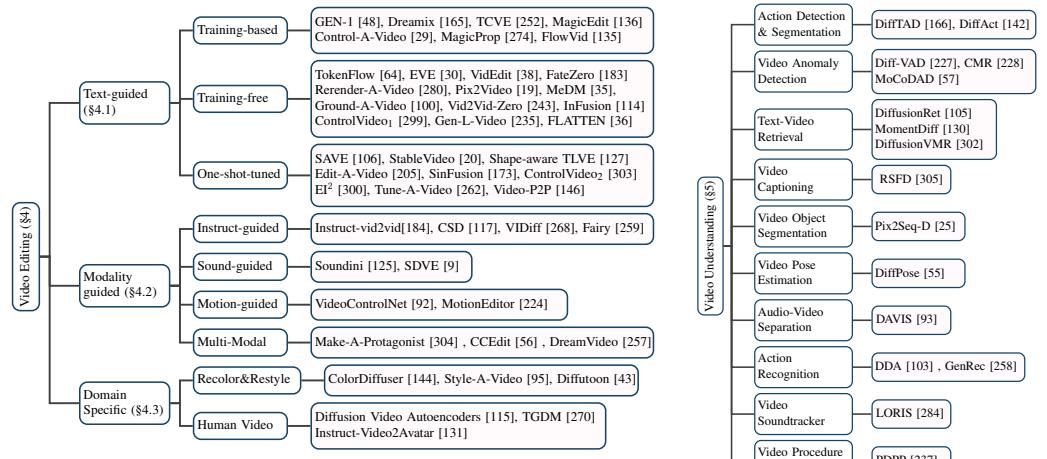


Fig. 5. Taxonomy of Video Editing. Key aspects of Video Editing include General Text-guided Video Editing, Modality-guided Video Editing and Domain-specific Video Editing.

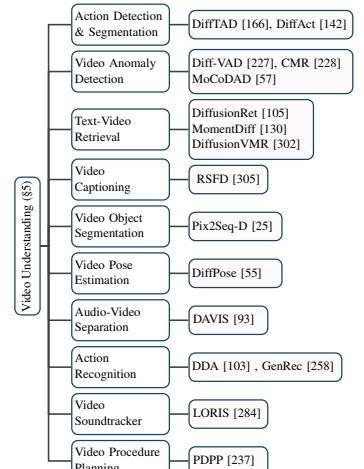


Fig. 6. Taxonomy of diffusion-based Video Understanding.

4 Video Editing

With the development of diffusion models, there has been an exponential growth in the number of research studies in video editing. As a consensus of many researches [20, 114, 165, 303], video editing tasks should satisfy the following criteria: (1) fidelity: each frame should be consistent in content with the corresponding frame of the original video; (2) alignment: the output video should be aligned with the input control information; (3) quality: the generated video should be temporal consistent and in high quality. While a pre-trained image diffusion model can be utilized for video editing by processing frames individually, the lack of semantic consistency across frames renders editing a video frame by frame infeasible, making video editing a challenging task. In this section, we divide video editing into three categories: Text-guided video editing (Sec. 4.1), Modality-guided video editing (Sec. 4.2), and Domain-specific video editing (Sec. 4.3). The taxonomy details of video editing are summarized in Fig. 5.

4.1 Text-guided Video Editing

In text-guided video editing, the user provides an input video and a text prompt which describes the desired attributes of the resulting video. However, unlike image editing, text-guided video editing represents new challenges of frame consistency and temporal modeling. In general, there are two main ways for text-based video editing: (1) training a T2V diffusion model on a large-scale text-video pairs dataset and (2) extending the pre-trained T2I diffusion models for video editing. The latter garnered more interest due to the fact that large-scale text-video datasets are hard to acquire, and training a T2V model is computationally expensive. To capture motion in videos, various temporal modules are introduced to T2I models. Nonetheless, methods inflating T2I models suffer from two critical issues: *Temporal inconsistency*, where the edited video exhibits flickering in vision across frame, and *Semantic disparity*, where videos are not altered in accordance with the semantics of given text prompts. Several studies address the problems from different perspectives.

4.1.1 Training-based Methods. The training-based approach refers to the method of training on a large-scale video-text dataset, enabling it to serve as a general video editing model.

GEN-1 [48] proposes a structure and content-aware model that fully controls temporal, content, and structural consistency. This model introduces temporal layers into a pre-trained T2I model and trains it jointly on images and videos, achieving real-time control over temporal consistency.

Dreamix [165] proposes two innovations: starting generation with a low-resolution version of the original video and fine-tuning the model on the original video. Additionally, they propose mixed fine-tuning with full temporal attention and temporal attention masking, enhancing motion editability.

TCVE [252] introduces a cohesive spatial-temporal modeling unit to connect the temporal U-Net and the pre-trained T2I U-Net, which effectively captures the temporal coherence of input videos.

Control-A-Video [29] is based on a pre-trained T2I diffusion model, incorporating a spatio-temporal self-attention module and trainable temporal layers. Additionally, they propose a first-frame conditioning strategy, allowing it to produce videos of any length using an auto-regressive method.

MagicEdit [136] separates the learning of content, structure, and motion in different frameworks, achieving high fidelity and temporal coherence.

MagicProp [274] improves video editing by separating appearance editing from motion-aware propagation. It edits a reference frame and then uses a diffusion model to generate each frame based on the previous frame, target depth, and reference appearance.

Unlike previous methods use flow as hard constraints, FlowVid[135] considers the potential imperfections in flow estimation. In this way, they include depth map as additional spatial condition, along with the temporal flow condition, enabling consistent and flexible video editing.

4.1.2 Training-free Methods. Training-free approach involves utilizing pre-trained T2I or T2V models and adapting them for video editing tasks in a zero-shot manner. Compared to training-based methods, training-free methods require no heavy training cost. However, they may suffer a few potential drawbacks. First of all, videos edited in a zero-shot manner may produce spatio-temporal distortion and inconsistency. Furthermore, methods utilizing T2V models might still incur high training and inference costs. We briefly examine the techniques used to address these issues.

TokenFlow [64] demonstrates that consistency in edited videos can be achieved by enforcing consistency in the diffusion feature space. Specifically, this is accomplished by sampling keyframes, jointly editing them, and propagating the features from the keyframes to all other frames based on the correspondences provided by the original video features. This process explicitly maintains consistency and a fine-grained shared representation of the original video features.

VidEdit [38] combines atlas-based [7] and pre-trained T2I [193] models, which not only exhibit high temporal consistency but also provide object-level control over video content appearance. The method involves decomposing videos into layered neural atlases with a semantically unified representation of content, and then applying a pre-trained, text-driven image diffusion model for zero-shot atlas editing. Concurrently, it preserves structure in atlas space by encoding both temporal appearance and spatial placement.

Rerender-A-Video [280] employs hierarchical cross-frame constraints to enforce temporal consistency. It uses optical flow to apply dense cross-frame constraints, with the previously rendered frame as a low-level reference and the first rendered frame as an anchor to maintain consistency in style, shape, texture, and color.

To address the issues of heavy costs in atlas learning [7] and per-video tuning [262], FateZero stores comprehensive attention maps at every stage of the inversion process to maintain superior motion and structural information. Additionally, it incorporates spatial-temporal blocks to enhance visual consistency.

Vid2Vid-Zero [243] utilizes a null-text inversion [164] module to align text with video, a spatial regularization module for video-to-video fidelity, and a cross-frame modeling module for temporal consistency. Similar to FateZero [183], it also incorporates a spatial-temporal attention module.

Pix2Video [19] initially utilizes a pre-trained structure-guided T2I model to conduct text-guided edits on an anchor frame, ensuring the generated image remains true to the edit prompt. Subsequently, they progressively propagate alterations to future frames using self-attention feature injection, maintaining temporal coherence.

InFusion [114] consists of two components. First, it integrates features from the residual block in decoder layers and attention features into the denoising pipeline for the editing prompt, showcasing its zero-shot editing capability. Second, it merges attention for edited and unedited concepts using mask extraction from cross-attention maps, ensuring consistency.

ControlVideo₁ [299] directly adopts the weights from ControlNet [297], extending self-attention with fully cross-frame interaction to achieve high-quality and consistency. To manage long-video editing tasks, it implements a hierarchical sampler that divides the long video into short clips and attains global coherence by conditioning on pairs of keyframes.

EVE [30] proposes two strategies to reinforce temporal consistency: *Depth Map Guidance* to locate spatial layouts and motion trajectories of moving objects as well as *Frame-Align Attention* which forces the model to place attention on both previous and current frames.

MeDM [35] utilizes explicit optical flows to establish a pragmatic encoding of pixel correspondences across video frames, thus maintaining temporal consistency. Furthermore, they iteratively align noisy pixels across video frames using the provided temporal correspondence guidance derived from optical flows.

Gen-L-Video [235] addresses long video editing by treating them as overlapping short segments. Using Temporal Co-Denoising methods, it adapts existing short video editing models [19, 79, 262] to edit videos with hundreds of frames while preserving consistency.

FLATTEN [36] incorporates optical flow into the attention mechanism of the diffusion model. The proposed Flow-guided attention allows patches from different frames to align on the same flow path within the attention module, enabling mutual attention and enhancing video editing consistency.

4.1.3 One-shot-tuned Methods. One-shot tuned method entails fine-tuning a pre-trained T2I model using a specific video instance, enabling the generation of videos with similar motion or content. While it requires extra training expenses, these approaches provide greater editing flexibility compared to training-free methods.

SinFusion [173] pioneers one-shot-tuned diffusion-based models that learn motions from a single input video using only a few frames. Its backbone is a fully convolutional DDPM [82] network, allowing it to generate images of any size.

SAVE [106] finetunes the spectral shift of the parameter space such that the underlying motion concept as well as content information in the input video is learned. Also, it proposes a spectral shift regularizer to restrict the changes.

Edit-A-Video [205] contains two stages: the first stage inflates a pre-trained T2I model to the T2V model and finetunes it using a single <text, video> pair while the second stage is the conventional diffusion and denoising process. A key observation is that edited videos often suffer from background inconsistency. To address such an issue, they propose a masking method called *sparse-causal blending*, which automatically generates a mask to approximate the edited region.

Tune-A-Video [262] leverages a sparse spatio-temporal attention mechanism that only visits the first and the former video frames, together with an efficient tuning strategy that only updates the projection matrices in the attention blocks. Furthermore, it seeks structural guidance from input video at inference time to make up for the lack of motion consistency.

Instead of using a T2I model, Video-P2P [146] alters it into a Text-to-set model (T2S) by replacing self-attentions with frame-attentions, which yields a model that generates a set of semantically-consistent images. Furthermore, they use a decoupled-guidance strategy to improve the robustness to the change of prompts.

ControlVideo₂ [303] mainly focuses on improving attention modules in the diffusion model and ControlNet [297]. They transform the original spatial self-attention into key-frame attention, which aligns all frames with a selected one. Additionally, they incorporate temporal attention modules to preserve consistency.

Shape-aware TLVE [127] utilizes the T2I model and handles shape changes by propagating the deformation field between the input and edited keyframe to all frames.

EI² [300] makes two key innovations: the Shift-restricted Temporal Attention Module (STAM) to restrict newly introduced parameters in the Temporal Attention module, resolving the semantic disparity, as well as the Fine-coarse Frame Attention Module (FFAM) for temporal consistency, which leverages the information on the temporal dimension by sampling along the spatial dimension. Combining these techniques, they create a T2V diffusion model.

StableVideo [20] designs an inter-frame propagation mechanism on top of the existing T2I model and an aggregation network to generate the edited atlases from the key frames, thus achieving temporal and spatial consistency.

4.2 Other Modality-guided Video Editing

Most of the methods introduced previously focus on text-guided video editing. In this subsection, we will focus on video editing guided by other modalities (*e.g.*, Instruct and Sound).

4.2.1 Instruct-guided Video Editing. Instruct-guided video editing aims to generate video based on the given input video and instructions. Due to the lack of video-instruction datasets, InstructVid2Vid [184] leverages the combined use of ChatGPT, BLIP [129], and Tune-A-Video [262] to acquire input videos, instructions and edited videos triplets at a relatively low cost. During training, they propose the Frame Difference Loss, guiding the model to generate temporal consistent frames. CSD [117] first uses Stein variational gradient descent (SVGD), where multiple samples share their knowledge distilled from diffusion models to accomplish inter-sample consistency. Then, they combine Collaborative Score Distillation (CSD) with Instruct-Pix2Pix [14] to achieve coherent editing of multiple images with instruction. VIDiff [268] is based on a multimodal instruction-guided editing diffusion model, unifying tasks including video editing, video recoloring, video object segmentation, and low-level tasks. It demonstrates the tremendous potential of video diffusion models and the feasibility of applying a unified architecture to various tasks. Fairy[259] samples anchor frames where the features extracted from them can be propagated to frames with high similarity with them, allowing for parallel processing among different frame groups. Hence, it ensures temporal consistency by sharing global features as well as reduces the memory requirement.

4.2.2 Sound-guided Video Editing. The goal of sound-guided video editing is to make visual changes consistent with the sound in the targeted region. To achieve this goal, Soundini [125] presents local sound guidance and optical flow guidance for diffusion sampling. Specifically, the audio encoder makes sound latent representation semantically consistent with the latent image representation. Based on a diffusion model, SDVE [9] introduces a feature concatenation mechanism for temporal coherence. They further condition the network on speech by feeding spectral feature embeddings with the noise signal throughout the residual layers.

4.2.3 Motion-guided Video Editing. Inspired by the video coding process, VideoControlNet [92] combines a diffusion model with ControlNet [297]. It designates the first frame as the I-frame and divides the remaining frames into groups of pictures (GoP). The last frame of each GoP is set as the P-frame, while others are B-frames. For an input video, the model generates the I-frame using the diffusion model and ControlNet. P-frames are then generated through the motion-guided P-frame generation module (MgPG), leveraging optical flow information. Finally, B-frames are interpolated based on the reference I/P-frames and motion information, avoiding the time-consuming diffusion model. MotionEditor[224] features a two-branch architecture (*i.e.* a reconstruction branch and an editing branch), which complements ControlNet[297] by seamlessly enforcing temporal motion correspondence, enabling high-fidelity editing and temporal consistency.

4.2.4 Multi-Modal Video Editing. Make-A-Protagonist [304] introduces a multi-modal conditioned video editing framework to alter the protagonist. It employs BLIP-2 [129] for video captioning, CLIP Vision Model [186] and DALLE-2 Prior [188] for visual and textual clue encoding, and ControlNet [297] for video consistency. During inference, a mask-guided denoising sampling technique combines these elements to achieve annotation-free video editing.

CCEdit [56] decouples video structure and appearance for controllable and creative video editing. It preserves the video structure using the foundational ControlNet [297] while allowing appearance editing through text prompts, personalized model weights, and customized center frames.

DreamVideo[257] disentangles the personalized video editing task into two stages, subject learning and motion learning. They utilize a light-weight adapter to capture the appearance of given subject through texture inversion and another adapter to model target motion pattern, reducing the complexity of optimization and allowing for more flexible customization.

4.3 Domain-specific Video Editing

In this subsection, we will provide a brief overview of several video editing techniques tailored for specific domains, starting with video recoloring and video style transfer methods in Sec. 4.3.1, followed by several video editing methods designed for human-centric videos in Sec. 4.3.2.

4.3.1 Recolor & Restyle. • **Recolor** Video colorization aims to infer realistic and consistent colors for grayscale frames, balancing temporal, spatial, semantic consistency, color richness, and faithfulness. ColorDiffuser [144], built on a pre-trained T2I model, introduces Color Propagation Attention to replace optical flow and an Alternated Sampling Strategy to capture spatio-temporal relationships between adjacent frames.

• **Restyle** Style-A-Video [95] designs a combined way of control conditions: text for style guidance, video frames for content guidance, and attention maps for detail guidance. Notably, the work features zero-shot, namely, no additional per-video training or fine-tuning is required. Diffutoon[43] solve the toon shading task by dividing it into four sub-problems: stylization, consistency enhancement, structure guidance and colorization. They utilize Stable Diffusion[193] for stylization while ControlNet[297] is used for both outline-based generation and video coloring.

4.3.2 Human Video Editing. Diffusion Video Autoencoders [115] proposes a diffusion video autoencoder that extracts a single time-invariant feature (identity) and per-frame time-varient features (motion and background) from a given video and further manipulates the single invariant feature for the desired attribute, which enables temporal-consistent editing and efficient computing. Instruct-Video2Avatar [131] takes in a talking head video and an editing instruction and outputs an edited version of 3D neural head avatar. They simultaneously leverage [14] for image editing, EbSynth [99] for video stylization, and INSTA [312] for a photo-realistic 3D neural head avatar. TGDM [270] adopts the zero-shot CLIP-guided model to achieve flexible emotion control. Furthermore, they propose a pipeline based on the multi-conditional diffusion model to afford complex texture and identity transfer.

5 Video Understanding

In addition to its application in generative tasks, such as video generation and editing, diffusion model has also been explored in video understanding tasks such as video temporal segmentation [142, 166], video anomaly detection [227, 228], text-video retrieval [105, 130], etc., as will be introduced in this section. The taxonomy details of video understanding are summarized in Fig. 6.

• **Temporal Action Detection& Segmentation** Inspired by DiffusionDet [22], DiffTAD [166] explores the application of diffusion models to the task of temporal action detection. This involves diffusing ground truth proposals of long videos and subsequently learning the denoising process, which is done by introducing a specialized temporal location query within the DETR [16] architecture. Notably, the approach achieves state-of-the-art performance results on benchmarks such as ActivityNet [119] and THUMOS [97].

Similarly, DiffAct [142] addresses the task of temporal action segmentation using a comparable approach, where action segments are iteratively generated from random noise with input video features as conditions. The effectiveness of the proposed method is validated on widely-used benchmarks, including GTEA [51], 50Salads [216], and Breakfast [120].

• **Video Anomaly Detection** Dedicated to unsupervised video anomaly detection, Diff-VAD [227] and CMR [228] harness the reconstruction capability of the diffusion model to identify anomalous videos, as high reconstruction error typically indicates abnormality. Experiments conducted on two large-scale benchmarks [148, 218] demonstrate the effectiveness of such a paradigm, consequently significantly improving performance compared to prior research.

MoCoDAD [57] focuses on skeleton-based video anomaly detection. The method applies the diffusion model to generate diverse and plausible future motions based on past actions of individuals. By statistically aggregating future patterns, anomalies are detected when a generated set of actions deviates from actual future trends.

- **Text-Video Retrieval** DiffusionRet [105] approaches retrieval as generating a joint distribution $p(\text{candidates}, \text{query})$ from noise. It combines generative and contrastive losses to train the generator and feature extractor, respectively, merging generative and discriminative techniques. This approach excels in open-domain scenarios, showing strong generalization ability.

MomentDiff [130] and DiffusionVMR [302] tackle video moment retrieval by converting time intervals into random noise and learning to denoise them back to their original states. This method trains the model to map random positions to actual intervals, improving the accuracy of localizing video segments based on textual descriptions.

- **Video Captioning** RSFD [305] examines the frequently neglected long-tail problem in video captioning. It presents a new Refined Semantic enhancement approach for Frequency Diffusion (RSFD), which improves captioning by constantly recognizing the linguistic representation of infrequent tokens. This allows the model to comprehend the semantics of low-frequency tokens, resulting in enhanced caption generation.

- **Video Object Segmentation** Pix2Seq-D [25] redefines panoramic segmentation as a discrete data generation problem. It employs a diffusion model based on analog bits [26] to model panoptic masks, utilizing a versatile architecture and loss function. Furthermore, Pix2Seq-D [25] can model videos by incorporating predictions from previous frames, which enables the automatic learning of object instance tracking and video object segmentation.

- **Video Pose Estimation** DiffPose [55] tackles video-based human pose estimation by treating it as a conditional heatmap generation task. It uses a Spatio-Temporal representation learner to aggregate features across frames and a multi-scale feature interaction mechanism to refine keypoint representations by establishing correlations across different scales.

- **Audio-Video Separation** DAVIS [93] tackles the audio-visual sound source separation task using a generative approach. The model employs a diffusion process to generate separated magnitudes from Gaussian noise, conditioned on the audio mixture and visual content.

- **Action Recognition** DDA [103] enhances skeleton-based human action recognition by using diffusion-based data augmentation to generate high-quality, diverse action sequences. Experiments demonstrate the method's advantages in naturalness and diversity and its effectiveness when applied to existing action recognition models.

GenRec [258] explores applying spatiotemporal priors of video diffusion model to recognition tasks. It uses the pretrained SVD [10] as the backbone and adds a classification head on top of the generative model, enabling the model to support both generation and classification tasks. Experimental results demonstrate that the two tasks can complement each other, achieving strong performance.

- **Video SoundTracker** LORIS [284] focuses on generating music soundtracks that synchronize with rhythmic visual cues. The system utilizes a latent conditional diffusion probabilistic model for waveform synthesis. Moreover, it incorporates context-aware conditioning encoders to account for temporal information, facilitating long-term waveform generation. The authors have also broadened the applicability of the model to various sports scenarios and is capable of producing long-term soundtracks with exceptional musical quality and rhythmic correspondence.

- **Video Procedure Planning** PDPP [237] addresses procedure planning in instructional videos using a diffusion model to represent the distribution of the entire intermediate action sequence, transforming planning into a sampling process. The method employs a diffusion-based U-Net model for precise conditional guidance from initial and final observations, improving the learning and sampling of action sequences from the distribution.

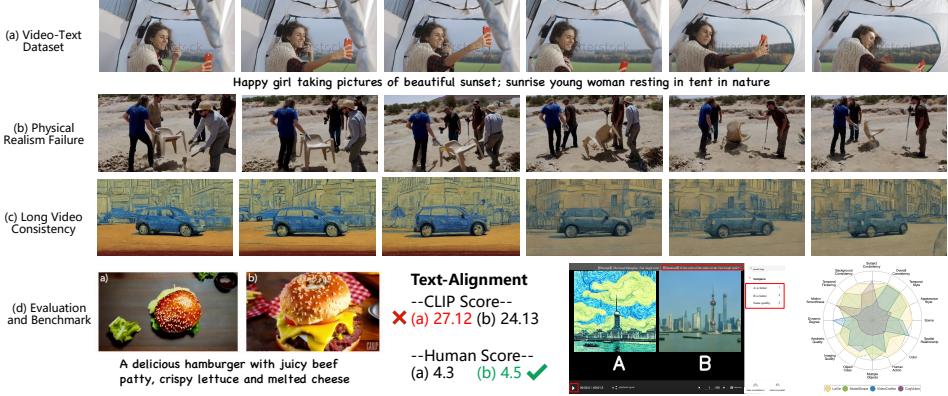


Fig. 7. Challenges and future trends. (a) Lack of high-quality video-text dataset [5], (b) Physical Realism failure [176], (c) Long video consistency failure [268], (d) Reliable Evaluation [261] and Benchmark [96].

6 Challenges and Future Trends

Despite the fact that diffusion-based methods have achieved significant advances in video generation, editing and understanding, there are still certain open problems worthy of exploration. In this section, we summarize the current challenges and potential future trends. We have also provided qualitative results in Fig. 7 to illustrate the current limitations and future trends.

- **Collecting Large-scale Video-Text Datasets** The substantial achievements in Text-to-Image synthesis primarily stemmed from the availability of billions of high-quality (text, image) pairs. However, the commonly used datasets for Text-to-Video (T2V) tasks are relatively small in scale and gathering equally extensive datasets for video content is a considerably challenging endeavor. For example, the WebVid dataset [5] contains only 10 million instances and has a significant drawback of its limited visual quality, with a low resolution of 360P, further compounded by watermark artifacts. Despite ongoing efforts to develop new methods for obtaining datasets [5, 28, 71, 244], there remains a pressing need to enhance dataset scale, annotation accuracy, and video quality. The successes of SVD [10] and Sora [176] demonstrate the effectiveness of scaling datasets. However, since they used private datasets, high-quality open-source datasets are crucial for video generation research.

- **Physical Realism and Long Video Generation** Even the most advanced video diffusion generation models, such as Sora [176], exhibit some limitations in accurately describing complex scenes, showing inconsistencies in the portrayal of physical principles in videos of complex scenes, such as incorrect simulations of the rigid structure of chairs and unrealistic physical interactions. Moreover, most video generation models [11, 207, 266] currently can only produce videos shorter than 10 seconds. Commonly used autoregressive methods [178] for generating long videos suffer from error accumulation, resulting in poorer quality in later frames. Besides, multi-stage coarse-to-fine methods [282] for long video generation can be complex and time-consuming. Future efforts should aim to explore more physically accurate and realistic video generation, as well as consistency and stability in long video synthesis.

- **Efficient Training and Inference** The heavy training cost associated with T2V models presents a significant challenge, with some tasks necessitating the use of hundreds of GPUs [11, 71]. Despite the efforts by methods [266] to reduce training cost, both the magnitude of dataset and temporal complexity remain a critical concern. With the rise of Sora [176], the high training cost of training from scratch has become a significant challenge in video generation. More efficient compression of video representations [238], exploration of effective spatiotemporal modeling methods [60, 266], and acceleration of training and inference times [247, 296] are important research directions.

• **Benchmark and Evaluation Methods** Although benchmarks [50, 96, 149, 150, 214, 271] and evaluation methods [186, 231] for open-domain video generation exist, they are relatively limited in scope, as is demonstrated in [33]. Due to the absence of ground truth for the generated videos in Text-to-Video (T2V) generation, existing metrics such as Fréchet Video Distance (FVD) [231] and Inception Score (IS) [200] primarily emphasize the disparities between generated and real video distributions. This makes it challenging to have a comprehensive evaluation metric that accurately reflects video generation quality. Currently, there is considerable reliance on user AB testing and subjective scoring, which is labor-intensive and potentially biased due to subjectivity. Constructing more tailored evaluation benchmarks [96, 149] and metrics [62, 116, 261] in the future is also a meaningful avenue of research.

• **More Controllable Video Editing** The task of video editing has evolved alongside the development of video generation. Although existing video editing models have achieved impressive results in video style transfer [64, 280], there are still limitations in certain tasks. For instance, previous video editing methods often exhibit noticeable temporal inconsistencies when controlling object replacement [262]. Most video editing methods rely on detailed text descriptions, which limits their control and generality [184, 268]. Additionally, there are limitations in current methods for multi-object editing, motion editing [224], and long video editing [235]. As base models for video generation continue to develop [10, 176], the future research trend will be towards more controllable video generation models with stronger generality and multi-modal capabilities.

7 Conclusion

This survey offered an in-depth exploration of the latest developments in the era of AIGC (AI-generated Content) with a focus on video diffusion models. To the best of our knowledge, this is the first work of its kind. We provided a comprehensive overview of the fundamental concepts of the diffusion process, popular benchmark datasets, and commonly used evaluation metrics. Building upon this foundation, we comprehensively reviewed over 100 different works focusing on the task of video generation, editing and understanding, and categorized them according to their technical perspectives and research objectives. Furthermore, in the experimental section, we meticulously described the experimental setups and conducted a fair comparative analysis across various benchmark datasets. In the end, we put forth several research directions for the future of video diffusion models.

Acknowledge This work was supported in part by National Natural Science Foundation of China (No. 62032006).

References

- [1] H. Alqahtani, M. Kavakli-Thorne, G. Kumar, and F. SBSSTC. An analysis of evaluation metrics of gans. In *ICITA*, 2019.
- [2] J. An, S. Zhang, H. Yang, S. Gupta, J.-B. Huang, J. Luo, and X. Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv:2304.08477*, 2023.
- [3] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell. Localizing moments in video with natural language. In *ICCV*, 2017.
- [4] O. Avrahami, D. Lischinski, and O. Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022.
- [5] M. Bain, A. Nagrani, G. Varol, and A. Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.
- [6] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv:2211.01324*, 2022.
- [7] O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel. Text2live: Text-driven layered image and video editing. In *ECCV*, 2022.
- [8] G. Batzolis, J. Stanczuk, C.-B. Schönlieb, and C. Etmann. Conditional image generation with score-based diffusion models. *arXiv:2111.13606*, 2021.
- [9] D. Bigioi, S. Basak, H. Jordan, R. McDonnell, and P. Corcoran. Speech driven video editing via an audio-conditioned diffusion model. *arXiv:2301.04474*, 2023.

- [10] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- [11] 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 *CVPR*, 2023.
- [12] S. Bond-Taylor, P. Hessey, H. Sasaki, T. P. Breckon, and C. G. Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *ECCV*, 2022.
- [13] T. Brooks, J. Hellsten, M. Aittala, T.-C. Wang, T. Aila, J. Lehtinen, M.-Y. Liu, A. Efros, and T. Karras. Generating long videos of dynamic scenes. In *NeurIPS*, 2022.
- [14] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.
- [15] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.
- [16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.
- [17] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman. A short note about kinetics-600. *arXiv:1808.01340*, 2018.
- [18] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.
- [19] D. Ceylan, C.-H. P. Huang, and N. J. Mitra. Pix2video: Video editing using image diffusion. In *ICCV*, 2023.
- [20] W. Chai, X. Guo, G. Wang, and Y. Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *ICCV*, 2023.
- [21] M. Chang, A. Prakash, and S. Gupta. Look ma, no hands! agent-environment factorization of egocentric videos. *arXiv:2305.16301*, 2023.
- [22] S. Chen, P. Sun, Y. Song, and P. Luo. Diffusiondet: Diffusion model for object detection. In *ICCV*, 2022.
- [23] S. Chen, M. Xu, J. Ren, Y. Cong, S. He, Y. Xie, A. Sinha, P. Luo, T. Xiang, and J.-M. Perez-Rua. Gentron: Diffusion transformers for image and video generation. In *CVPR*, 2024.
- [24] T. Chen and L. Li. Fit: Far-reaching interleaved transformers. *arXiv preprint arXiv:2305.12689*, 2023.
- [25] T. Chen, L. Li, S. Saxena, G. Hinton, and D. J. Fleet. A generalist framework for panoptic segmentation of images and videos. In *ICCV*, 2022.
- [26] T. Chen, R. Zhang, and G. Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *ICLR*, 2023.
- [27] T.-S. Chen, C. H. Lin, H.-Y. Tseng, T.-Y. Lin, and M.-H. Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv:2304.14404*, 2023.
- [28] 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 *CVPR*, 2024.
- [29] W. Chen, J. Wu, P. Xie, H. Wu, J. Li, X. Xia, X. Xiao, and L. Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv:2305.13840*, 2023.
- [30] Y. Chen, X. Dong, T. Gan, C. Zhou, M. Yang, and Q. Guo. Eve: Efficient zero-shot text-based video editing with depth map guidance and temporal consistency constraints. *arXiv:2308.10648*, 2023.
- [31] Z. Chen, J. Qing, and J. H. Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *arXiv:2305.11675*, 2023.
- [32] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.
- [33] I. Chivilieva, P. Lynch, T. E. Ward, and A. F. Smeaton. Measuring the quality of text-to-video model outputs: Metrics and dataset. *arXiv:2309.08009*, 2023.
- [34] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *ICCV*, 2021.
- [35] E. Chu, T. Huang, S.-Y. Lin, and J.-C. Chen. Medm: Mediating image diffusion models for video-to-video translation with temporal correspondence guidance. *arXiv preprint arXiv:2308.10079*, 2023.
- [36] Y. Cong, M. Xu, C. Simon, S. Chen, J. Ren, Y. Xie, J.-M. Perez-Rua, B. Rosenhahn, T. Xiang, and S. He. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv:2310.05922*, 2023.
- [37] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.
- [38] P. Couairon, C. Rambour, J.-E. Haugéard, and N. Thome. Videdit: Zero-shot and spatially aware text-driven video editing. *arXiv:2306.08707*, 2023.

- [39] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018.
- [40] D. Danier, F. Zhang, and D. Bull. Ldmvfi: Video frame interpolation with latent diffusion models. *arXiv preprint arXiv:2303.09508*, 2023.
- [41] M. Ding, W. Zheng, W. Hong, and J. Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. In *NeurIPS*, 2022.
- [42] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. Carla: An open urban driving simulator. In *CoRL*, 2017.
- [43] Z. Duan, C. Wang, C. Chen, W. Qian, and J. Huang. Diffutoon: High-resolution editable toon shading via diffusion models. *arXiv preprint arXiv:2401.16224*, 2024.
- [44] Z. Duan, L. You, C. Wang, C. Chen, Z. Wu, W. Qian, J. Huang, F. Chao, and R. Ji. Diffsynth: Latent in-iteration deflickering for realistic video synthesis. *arXiv:2308.03463*, 2023.
- [45] F. Ebert, C. Finn, A. X. Lee, and S. Levine. Self-supervised visual planning with temporal skip connections. *CoRL*, 2017.
- [46] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *Robotics: Science and Systems*, 2022.
- [47] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang. Clap learning audio concepts from natural language supervision. In *ICASSP*, 2023.
- [48] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, 2023.
- [49] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.
- [50] F. Fan, C. Luo, J. Zhan, and W. Gao. Aigcbench: Comprehensive evaluation of image-to-video content generated by ai. *arXiv preprint arXiv:2401.01651*, 2024.
- [51] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011.
- [52] H. Fei, S. Wu, W. Ji, H. Zhang, and T.-S. Chua. Empowering dynamics-aware text-to-video diffusion with large language models. *arXiv:2308.13812*, 2023.
- [53] H. Fei, S. Wu, H. Zhang, T.-S. Chua, and S. Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing, 2024.
- [54] Q. Feng, Z. Xing, Z. Wu, and Y.-G. Jiang. Fdgaussian: Fast gaussian splatting from single image via geometric-aware diffusion model. *arXiv preprint arXiv:2403.10242*, 2024.
- [55] R. Feng, Y. Gao, T. H. E. Tse, X. Ma, and H. J. Chang. Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In *ICCV*, 2023.
- [56] R. Feng, W. Weng, Y. Wang, Y. Yuan, J. Bao, C. Luo, Z. Chen, and B. Guo. Credit: Creative and controllable video editing via diffusion models. *arXiv:2309.16496*, 2023.
- [57] A. Flaborea, L. Collorone, G. D'Amely, S. D'Arrigo, B. Prenkaj, and F. Galasso. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In *ICCV*, 2023.
- [58] T.-J. Fu, L. Yu, N. Zhang, C.-Y. Fu, J.-C. Su, W. Y. Wang, and S. Bell. Tell me what happened: Unifying text-guided video completion via multimodal masked video generation. In *CVPR*, 2023.
- [59] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, J. Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Found. Trends Comput. Graph. Vis.*, 2022.
- [60] 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.
- [61] S. Ge, T. Hayes, H. Yang, X. Yin, G. Pang, D. Jacobs, J.-B. Huang, and D. Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *ECCV*, 2022.
- [62] S. Ge, A. Mahapatra, G. Parmar, J.-Y. Zhu, and J.-B. Huang. On the content bias in fréchet video distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7277–7288, 2024.
- [63] S. Ge, S. Nah, G. Liu, T. Poon, A. Tao, B. Catanzaro, D. Jacobs, J.-B. Huang, M.-Y. Liu, and Y. Balaji. Preserve your own correlation: A noise prior for video diffusion models. *arXiv:2305.10474*, 2023.
- [64] M. Geyer, O. Bar-Tal, S. Bagon, and T. Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv:2307.10373*, 2023.
- [65] R. Girdhar, M. Singh, A. Brown, Q. Duval, S. Azadi, S. S. Ramabhatla, A. Shah, X. Yin, D. Parikh, and I. Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.
- [66] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014.
- [67] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 2020.
- [68] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al. The " something something" video database for learning and evaluating visual common sense.

In *ICCV*, 2017.

- [69] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.
- [70] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J Mach Learn Res*, 2012.
- [71] 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. *arXiv:2309.03549*, 2023.
- [72] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022.
- [73] X. Gu, C. Wen, J. Song, and Y. Gao. Seer: Language instructed video prediction with latent diffusion models. *arXiv:2303.14897*, 2023.
- [74] Z. Gu, H. Chen, Z. Xu, J. Lan, C. Meng, and W. Wang. Diffusioninst: Diffusion model for instance segmentation. *arXiv:2212.02773*, 2022.
- [75] Y. Guo, C. Yang, A. Rao, Y. Wang, Y. Qiao, D. Lin, and B. Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv:2307.04725*, 2023.
- [76] A. Gupta, L. Yu, K. Sohn, X. Gu, M. Hahn, L. Fei-Fei, I. Essa, L. Jiang, and J. Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023.
- [77] W. Harvey, S. Naderiparizi, V. Masrani, C. Weilbach, and F. Wood. Flexible diffusion modeling of long videos. In *NeurIPS*, 2022.
- [78] Y. He, M. Xia, H. Chen, X. Cun, Y. Gong, J. Xing, Y. Zhang, X. Wang, C. Weng, Y. Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv:2307.06940*, 2023.
- [79] Y. He, T. Yang, Y. Zhang, Y. Shan, and Q. Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv:2211.13221*, 2022.
- [80] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023.
- [81] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022.
- [82] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.
- [83] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 2022.
- [84] J. Ho and T. Salimans. Classifier-free diffusion guidance. In *NeurIPS*, 2021.
- [85] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. In *NeurIPS*, 2022.
- [86] S. Hong, J. Seo, S. Hong, H. Shin, and S. Kim. Large language models are frame-level directors for zero-shot text-to-video generation. *arXiv:2305.14330*, 2023.
- [87] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023.
- [88] T. Höppé, A. Mehrjou, S. Bauer, D. Nielsen, and A. Dittadi. Diffusion models for video prediction and infilling. *Trans. Mach. Learn. Res.*, 2022.
- [89] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.
- [90] L. Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, pages 8153–8163, 2024.
- [91] Y. Hu, Z. Chen, and C. Luo. Lamd: Latent motion diffusion for video generation. *arXiv:2304.11603*, 2023.
- [92] Z. Hu and D. Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv:2307.14073*, 2023.
- [93] C. Huang, S. Liang, Y. Tian, A. Kumar, and C. Xu. Davis: High-quality audio-visual separation with generative diffusion models. *arXiv:2308.00122*, 2023.
- [94] H. Huang, Y. Feng, C. Shi, L. Xu, J. Yu, and S. Yang. Free-bloom: Zero-shot text-to-video generator with ldm director and ldm animator. In *NeurIPS*, 2023.
- [95] N. Huang, Y. Zhang, and W. Dong. Style-a-video: Agile diffusion for arbitrary text-based video style transfer. *arXiv:2305.05464*, 2023.
- [96] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.
- [97] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. The thumos challenge on action recognition for videos “in the wild”. *CVIU*, 2017.
- [98] Y. Jafarian and H. S. Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *CVPR*, 2021.

- [99] O. Jamriska. Ebsynth: Fast example-based image synthesis and style transfer, 2018.
- [100] H. Jeong and J. C. Ye. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. *arXiv:2310.01107*, 2023.
- [101] Y. Jeong, W. Ryoo, S. Lee, D. Seo, W. Byeon, S. Kim, and J. Kim. The power of sound (tpos): Audio reactive video generation with stable diffusion. In *ICCV*, 2023.
- [102] Y. Ji, Z. Chen, E. Xie, L. Hong, X. Liu, Z. Liu, T. Lu, Z. Li, and P. Luo. Ddp: Diffusion model for dense visual prediction. *arXiv:2303.17559*, 2023.
- [103] Y. Jiang, H. Chen, and H. Ko. Spatial-temporal transformer-guided diffusion based data augmentation for efficient skeleton-based action recognition. *arXiv:2302.13434*, 2023.
- [104] Y. Jiang, S. Yang, T. L. Koh, W. Wu, C. C. Loy, and Z. Liu. Text2performer: Text-driven human video generation. In *ICCV*, 2023.
- [105] P. Jin, H. Li, Z. Cheng, K. Li, X. Ji, C. Liu, L. Yuan, and J. Chen. Diffusionret: Generative text-video retrieval with diffusion model. In *ICCV*, 2023.
- [106] N. Karim, U. Khalid, M. Joneidi, C. Chen, and N. Rahnavard. Save: Spectral-shift-aware adaptation of image diffusion models for text-guided video editing. *arXiv:2305.18670*, 2023.
- [107] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv:2304.06025*, 2023.
- [108] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.
- [109] T. Karras, M. Aittala, S. Laine, E. Häkkinen, J. Hellsten, J. Lehtinen, and T. Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021.
- [110] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.
- [111] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020.
- [112] B. Kawar, M. Elad, S. Ermon, and J. Song. Denoising diffusion restoration models. In *NeurIPS*, 2022.
- [113] 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 *ICCV*, 2023.
- [114] A. Khandelwal. Infusion: Inject and attention fusion for multi concept zero shot text based video editing. In *ICCVW*, 2023.
- [115] 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 *CVPR*, 2023.
- [116] P. J. Kim, S. Kim, and J. Yoo. Stream: Spatio-temporal evaluation and analysis metric for video generative models. *arXiv preprint arXiv:2403.09669*, 2024.
- [117] S. Kim, K. Lee, J. S. Choi, J. Jeong, K. Sohn, and J. Shin. Collaborative score distillation for consistent visual synthesis. In *NeurIPS*, 2023.
- [118] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *ICCV*, 2023.
- [119] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017.
- [120] H. Kuehne, A. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, 2014.
- [121] D. Lab. Deepfloyd if, 2023.
- [122] A. Lapid, I. Achituve, L. Bracha, and E. Fetaya. Gd-vdm: Generated depth for better diffusion-based video generation. *arXiv:2306.11173*, 2023.
- [123] G. Le Moing, J. Ponce, and C. Schmid. Ccvs: context-aware controllable video synthesis. In *NeurIPS*, 2021.
- [124] S. Lee, C. Kong, D. Jeon, and N. Kwak. Adiff: Audio-aligned video synthesis with text-to-image diffusion. In *CVPRW*, 2023.
- [125] S. H. Lee, S. Kim, I. Yoo, F. Yang, D. Cho, Y. Kim, H. Chang, J. Kim, and S. Kim. Soundini: Sound-guided diffusion for natural video editing. *arXiv:2304.06818*, 2023.
- [126] T. Lee, S. Kwon, and T. Kim. Grid diffusion models for text-to-video generation. In *CVPR*, 2024.
- [127] Y.-C. Lee, J.-Z. G. Jang, Y.-T. Chen, E. Qiu, and J.-B. Huang. Shape-aware text-driven layered video editing. In *CVPR*, 2023.
- [128] C. Li, C. Zhang, A. Waghwase, L.-H. Lee, F. Rameau, Y. Yang, S.-H. Bae, and C. S. Hong. Generative ai meets 3d: A survey on text-to-3d in aigc era. *arXiv:2305.06131*, 2023.
- [129] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

- [130] P. Li, C.-W. Xie, H. Xie, L. Zhao, L. Zhang, Y. Zheng, D. Zhao, and Y. Zhang. Momentdiff: Generative video moment retrieval from random to real. In *NeurIPS*, 2023.
- [131] S. Li. Instruct-video2avatar: Video-to-avatar generation with instructions. *arXiv:2306.02903*, 2023.
- [132] X. Li, W. Chu, Y. Wu, W. Yuan, F. Liu, Q. Zhang, F. Li, H. Feng, E. Ding, and J. Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023.
- [133] Z. Li, R. Tucker, N. Snavely, and A. Holynski. Generative image dynamics. *arXiv:2309.07906*, 2023.
- [134] L. Lian, B. Shi, A. Yala, T. Darrell, and B. Li. Llm-grounded video diffusion models. *arXiv:2309.17444*, 2023.
- [135] F. Liang, B. Wu, J. Wang, L. Yu, K. Li, Y. Zhao, I. Misra, J.-B. Huang, P. Zhang, P. Vajda, et al. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *CVPR*, 2024.
- [136] J. H. Liew, H. Yan, J. Zhang, Z. Xu, and J. Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv:2308.14749*, 2023.
- [137] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023.
- [138] H. Lin, A. Zala, J. Cho, and M. Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv:2309.15091*, 2023.
- [139] J. Lin, C. Gan, and S. Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019.
- [140] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [141] B. Liu, X. Liu, A. Dai, Z. Zeng, Z. Cui, and J. Yang. Dual-stream diffusion net for text-to-video generation. *arXiv:2308.08316*, 2023.
- [142] D. Liu, Q. Li, A. Dinh, T. Jiang, M. Shah, and C. Xu. Diffusion action segmentation. In *ICCV*, 2023.
- [143] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *NeurIPS*, 2024.
- [144] H. Liu, M. Xie, J. Xing, C. Li, and T.-T. Wong. Video colorization with pre-trained text-to-image diffusion models. *arXiv:2306.01732*, 2023.
- [145] J. Liu, W. Wang, W. Liu, Q. He, and J. Liu. Ed-t2v: An efficient training framework for diffusion-based text-to-video generation. In *IJCNN*, 2023.
- [146] S. Liu, Y. Zhang, W. Li, Z. Lin, and J. Jia. Video-p2p: Video editing with cross-attention control. *arXiv:2303.04761*, 2023.
- [147] V. Liu, T. Long, N. Raw, and L. Chilton. Generative disco: Text-to-video generation for music visualization. *arXiv:2304.08551*, 2023.
- [148] W. Liu, W. Luo, D. Lian, and S. Gao. Future frame prediction for anomaly detection—a new baseline. In *CVPR*, 2018.
- [149] 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 *CVPR*, 2024.
- [150] Y. Liu, L. Li, S. Ren, R. Gao, S. Li, S. Chen, X. Sun, and L. Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *NeurIPS*, 2024.
- [151] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.
- [152] H. Lu, G. Yang, N. Fei, Y. Huo, Z. Lu, P. Luo, and M. Ding. Vdt: An empirical study on video diffusion with transformers. *arXiv:2305.13311*, 2023.
- [153] S. Luo and W. Hu. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, 2021.
- [154] Z. Luo, D. Chen, Y. Zhang, Y. Huang, L. Wang, Y. Shen, D. Zhao, J. Zhou, and T. Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 2023.
- [155] 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. *arXiv preprint arXiv:2401.03048*, 2024.
- [156] Y. Ma, Y. He, X. Cun, X. Wang, Y. Shan, X. Li, and Q. Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv:2304.01186*, 2023.
- [157] K. Mei and V. Patel. Vidm: Video implicit diffusion models. In *AAAI*, 2023.
- [158] W. Menapace, A. Siarohin, I. Skorokhodov, E. Deyneca, T.-S. Chen, A. Kag, Y. Fang, A. Stolar, E. Ricci, J. Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *CVPR*, pages 7038–7048, 2024.
- [159] C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, and T. Salimans. On distillation of guided diffusion models. In *CVPR*, 2023.
- [160] C. Meng, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. In *ICLR*, 2022.
- [161] Midjourney. Midjourney., 2022.
- [162] 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 *ICCV*, 2019.

- [163] S. Mo and Y. Tian. Scaling diffusion mamba with bidirectional ssms for efficient image and video generation. *arXiv preprint arXiv:2405.15881*, 2024.
- [164] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2022.
- [165] E. Molad, E. Horwitz, D. Valevski, A. R. Acha, Y. Matias, Y. Pritch, Y. Leviathan, and Y. Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv:2302.01329*, 2023.
- [166] S. Nag, X. Zhu, J. Deng, Y.-Z. Song, and T. Xiang. Difftad: Temporal action detection with proposal denoising diffusion. In *ICCV*, 2023.
- [167] A. Nagrani, P. H. Seo, B. Seybold, A. Hauth, S. Manen, C. Sun, and C. Schmid. Learning audio-video modalities from image captions. In *ECCV*, 2022.
- [168] J. Nam, G. Lee, S. Kim, H. Kim, H. Cho, S. Kim, and S. Kim. Diffmatch: Diffusion model for dense matching. *arXiv:2305.19094*, 2023.
- [169] 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 *CVPR*, 2024.
- [170] H. Ni, C. Shi, K. Li, S. X. Huang, and M. R. Min. Conditional image-to-video generation with latent flow diffusion models. In *CVPR*, 2023.
- [171] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022.
- [172] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021.
- [173] Y. Nikankin, N. Haim, and M. Irani. Sinfusion: Training diffusion models on a single image or video. In *ICML*, 2022.
- [174] OpenAI. Chatgpt: A large-scale generative model for conversational ai, 2022.
- [175] OpenAI. Gpt-4 technical report. *arXiv:2303.08774*, 2023.
- [176] OpenAI. Sora, 2024.
- [177] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.
- [178] Y. Ouyang, H. Zhao, G. Wang, et al. Flexifilm: Long video generation with flexible conditions. *arXiv preprint arXiv:2404.18620*, 2024.
- [179] G. Parmar, R. Zhang, and J.-Y. Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, pages 11410–11420, 2022.
- [180] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.
- [181] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.
- [182] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
- [183] 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 *ICCV*, 2023.
- [184] B. Qin, J. Li, S. Tang, T.-S. Chua, and Y. Zhuang. Instructvid2vid: Controllable video editing with natural language instructions. *arXiv:2305.12328*, 2023.
- [185] 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. *arXiv:2308.07749*, 2023.
- [186] 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 *ICML*, 2021.
- [187] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 2021.
- [188] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022.
- [189] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021.
- [190] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020.
- [191] H. Reynaud, M. Qiao, M. Dombrowski, T. Day, R. Razavi, A. Gomez, P. Leeson, and B. Kainz. Feature-conditioned cascaded video diffusion models for precise echocardiogram synthesis. In *MICCAI*, 2023.
- [192] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele. Movie description. *IJCV*, 2017.
- [193] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

- [194] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.
- [195] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.
- [196] L. Ruan, Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, and B. Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *CVPR*, 2023.
- [197] 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 *CVPR*, 2023.
- [198] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.
- [199] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. In *TPAMI*, 2022.
- [200] M. Saito, S. Saito, M. Koyama, and S. Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. In *IJCV*, 2020.
- [201] T. Salimans and J. Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.
- [202] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze. How2: a large-scale dataset for multimodal language understanding. *arXiv:1811.00347*, 2018.
- [203] A. Sauer, K. Schwarz, and A. Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH*, 2022.
- [204] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv:2111.02114*, 2021.
- [205] C. Shin, H. Kim, C. H. Lee, S.-g. Lee, and S. Yoon. Edit-a-video: Single video editing with object-aware consistency. *arXiv:2303.07945*, 2023.
- [206] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. First order motion model for image animation. In *NeurIPS*, 2019.
- [207] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023.
- [208] I. Skorokhodov, S. Tulyakov, and M. Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*, 2022.
- [209] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.
- [210] Y. Song, C. Durkan, I. Murray, and S. Ermon. Maximum likelihood training of score-based diffusion models. In *NeurIPS*, 2021.
- [211] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019.
- [212] Y. Song and S. Ermon. Improved techniques for training score-based generative models. In *NeurIPS*, 2020.
- [213] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.
- [214] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [215] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015.
- [216] S. Stein and S. J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *ACMSE*, 2013.
- [217] J. C. Stroud, Z. Lu, C. Sun, J. Deng, R. Sukthankar, C. Schmid, and D. A. Ross. Learning video representations from textual web supervision. *arXiv:2007.14937*, 2020.
- [218] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, 2018.
- [219] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.
- [220] Z. Tang, Z. Yang, C. Zhu, M. Zeng, and M. Bansal. Any-to-any generation via composable diffusion. In *NeurIPS*, 2023.
- [221] L. Tian, Q. Wang, B. Zhang, and L. Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024.
- [222] Y. Tian, J. Ren, M. Chai, K. Olszewski, X. Peng, D. N. Metaxas, and S. Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *ICLR*, 2021.
- [223] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

- [224] S. Tu, Q. Dai, Z.-Q. Cheng, H. Hu, X. Han, Z. Wu, and Y.-G. Jiang. Motioneditor: Editing video motion via content-aware diffusion. In *CVPR*, 2024.
- [225] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 2018.
- [226] N. Tumanyan, M. Geyer, S. Bagor, and T. Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023.
- [227] A. O. Tur, N. Dall'Asen, C. Beyan, and E. Ricci. Exploring diffusion models for unsupervised video anomaly detection. *IEEE VCIP*, 2023.
- [228] A. O. Tur, N. Dall'Asen, C. Beyan, and E. Ricci. Unsupervised video anomaly detection with diffusion models conditioned on compact motion representations. 2023.
- [229] A. Ulhaq, N. Akhtar, and G. Pogrebna. Efficient diffusion models for vision: A survey. *arXiv preprint arXiv:2210.09292*, 2022.
- [230] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv:1812.01717*, 2018.
- [231] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Fvd: A new metric for video generation. In *ICLR*, 2019.
- [232] A. Vahdat, K. Kreis, and J. Kautz. Score-based generative modeling in latent space. In *NeurIPS*, 2021.
- [233] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 2017.
- [234] V. Voleti, A. Jolicœur-Martineau, and C. Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. In *NeurIPS*, 2022.
- [235] 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. *arXiv preprint arXiv:2305.18264*, 2023.
- [236] H. Wang, J. Cao, R. M. Anwer, J. Xie, F. S. Khan, and Y. Pang. Dformer: Diffusion-guided transformer for universal image segmentation. *arXiv:2306.03437*, 2023.
- [237] H. Wang, Y. Wu, S. Guo, and L. Wang. Pdpp: Projected diffusion for procedure planning in instructional videos. In *CVPR*, 2023.
- [238] J. Wang, Y. Jiang, Z. Yuan, B. Peng, Z. Wu, and Y.-G. Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *arXiv preprint arXiv:2406.09399*, 2024.
- [239] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2020.
- [240] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang. Modelscope text-to-video technical report. *arXiv:2308.06571*, 2023.
- [241] T. Wang, L. Li, K. Lin, C.-C. Lin, Z. Yang, H. Zhang, Z. Liu, and L. Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv:2307.00040*, 2023.
- [242] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018.
- [243] W. Wang, K. Xie, Z. Liu, H. Chen, Y. Cao, X. Wang, and C. Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv:2303.17599*, 2023.
- [244] W. Wang, H. Yang, Z. Tuo, H. He, J. Zhu, J. Fu, and J. Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv:2305.10874*, 2023.
- [245] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019.
- [246] 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. *arXiv:2306.02018*, 2023.
- [247] X. Wang, S. Zhang, H. Zhang, Y. Liu, Y. Zhang, C. Gao, and N. Sang. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023.
- [248] Y. Wang, J. Bao, W. Weng, R. Feng, D. Yin, T. Yang, J. Zhang, Q. Dai, Z. Zhao, C. Wang, et al. Microcinema: A divide-and-conquer approach for text-to-video generation. In *CVPR*, 2024.
- [249] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv:2309.15103*, 2023.
- [250] Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Chen, Y. Wang, P. Luo, Z. Liu, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv:2307.06942*, 2023.
- [251] Y. Wang, L. Jiang, and C. C. Loy. Styleinv: A temporal style modulated inversion network for unconditional video generation. In *ICCV*, 2023.
- [252] Y. Wang, Y. Li, X. Liu, A. Dai, A. Chan, and Z. Cui. Edit temporal-consistent videos with image diffusion model. *arXiv:2308.09091*, 2023.

- [253] Y. Wang, X. Ma, X. Chen, A. Dantcheva, B. Dai, and Y. Qiao. Leo: Generative latent image animator for human video synthesis. *arXiv:2305.03989*, 2023.
- [254] Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Process Mag*, 2009.
- [255] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004.
- [256] Z. Wang, Z. Yuan, X. Wang, T. Chen, M. Xia, P. Luo, and Y. Shan. Motionctrl: A unified and flexible motion controller for video generation. *arXiv preprint arXiv:2312.03641*, 2023.
- [257] 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 *CVPR*, 2024.
- [258] Z. Weng, X. Yang, Z. Xing, Z. Wu, and Y.-G. Jiang. Genrec: Unifying video generation and recognition with diffusion models. *arXiv preprint arXiv:2408.15241*, 2024.
- [259] B. Wu, C.-Y. Chuang, X. Wang, Y. Jia, K. Krishnakumar, T. Xiao, F. Liang, L. Yu, and P. Vajda. Fairy: Fast parallelized instruction-guided video-to-video synthesis. In *CVPR*, 2024.
- [260] J. Wu, W. Gan, Z. Chen, S. Wan, and H. Lin. Ai-generated content (aigc): A survey. *arXiv:2304.06632*, 2023.
- [261] J. Z. Wu, G. Fang, H. Wu, X. Wang, Y. Ge, X. Cun, D. J. Zhang, J.-W. Liu, Y. Gu, R. Zhao, et al. Towards a better metric for text-to-video generation. *arXiv preprint arXiv:2401.07781*, 2024.
- [262] J. Z. Wu, Y. Ge, X. Wang, W. Lei, Y. Gu, 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 *ICCV*, 2023.
- [263] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua. Next-gpt: Any-to-any multimodal llm. *arXiv:2309.05519*, 2023.
- [264] J. Xing, M. Xia, Y. Liu, Y. Zhang, Y. He, H. Liu, H. Chen, X. Cun, X. Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *arXiv:2306.00943*, 2023.
- [265] J. Xing, M. Xia, Y. Zhang, H. Chen, X. Wang, T.-T. Wong, and Y. Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023.
- [266] Z. Xing, Q. Dai, H. Hu, Z. Wu, and Y.-G. Jiang. Simda: Simple diffusion adapter for efficient video generation. In *CVPR*, 2024.
- [267] Z. Xing, Q. Dai, Z. Weng, Z. Wu, and Y.-G. Jiang. Aid: Adapting image2video diffusion models for instruction-guided video prediction. *arXiv preprint arXiv:2406.06465*, 2024.
- [268] Z. Xing, Q. Dai, Z. Zhang, H. Zhang, H. Hu, Z. Wu, and Y.-G. Jiang. Vidiff: Translating videos via multi-modal instructions with diffusion models. *arXiv preprint arXiv:2311.18837*, 2023.
- [269] 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 *CVPR*, 2018.
- [270] C. Xu, S. Zhu, J. Zhu, T. Huang, J. Zhang, Y. Tai, and Y. Liu. Multimodal-driven talking face generation via a unified diffusion-based generator. *arXiv:2305.02594*, 2023.
- [271] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016.
- [272] 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 *CVPR*, 2024.
- [273] 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 *CVPR*, 2022.
- [274] H. Yan, J. H. Liew, L. Mai, S. Lin, and J. Feng. Magicprop: Diffusion-based video editing via motion-aware appearance propagation. *arXiv preprint arXiv:2309.00908*, 2023.
- [275] W. Yan, Y. Zhang, P. Abbeel, and A. Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.
- [276] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput Surv*, 2022.
- [277] M. Yang, Y. Du, B. Dai, D. Schuurmans, J. B. Tenenbaum, and P. Abbeel. Probabilistic adaptation of text-to-video models. *arXiv:2306.01872*, 2023.
- [278] R. Yang, P. Srivastava, and S. Mandt. Diffusion probabilistic modeling for video generation. *arXiv:2203.09481*, 2022.
- [279] S. Yang, L. Zhang, Y. Liu, Z. Jiang, and Y. He. Video diffusion models with local-global context guidance. In *IJCAI*, 2023.
- [280] S. Yang, Y. Zhou, Z. Liu, and C. C. Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia*, 2023.
- [281] 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. *arXiv:2308.08089*, 2023.
- [282] S. Yin, C. Wu, H. Yang, J. Wang, X. Wang, M. Ni, Z. Yang, L. Li, S. Liu, F. Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv:2303.12346*, 2023.

- [283] YouTube. Youtube.
- [284] J. Yu, Y. Wang, X. Chen, X. Sun, and Y. Qiao. Long-term rhythmic video soundtracker. In *ICML*, 2023.
- [285] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*, 2022.
- [286] J. Yu, H. Zhu, L. Jiang, C. C. Loy, W. Cai, and W. Wu. CelebV-text: A large-scale facial text-video dataset. In *CVPR*, 2023.
- [287] S. Yu, W. Nie, D.-A. Huang, B. Li, J. Shin, and A. Anandkumar. Efficient video diffusion models via content-frame motion-latent decomposition. *arXiv preprint arXiv:2403.14148*, 2024.
- [288] S. Yu, K. Sohn, S. Kim, and J. Shin. Video probabilistic diffusion models in projected latent space. In *CVPR*, 2023.
- [289] S. Yu, J. Tack, S. Mo, H. Kim, J. Kim, J.-W. Ha, and J. Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *ICLR*, 2022.
- [290] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi. Merlot: Multimodal neural script knowledge models. In *NeurIPS*, 2021.
- [291] Y. Zeng, G. Wei, J. Zheng, J. Zou, Y. Wei, Y. Zhang, and H. Li. Make pixels dance: High-dynamic video generation. *arXiv preprint arXiv:2311.10982*, 2023.
- [292] F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, L. Liu, A. Kortylewski, C. Theobalt, and E. Xing. Multimodal image synthesis and editing: A survey and taxonomy. *TPAMI*, 2023.
- [293] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.
- [294] C. Zhang, C. Zhang, S. Zheng, Y. Qiao, C. Li, M. Zhang, S. K. Dam, C. M. Thwal, Y. L. Tun, L. L. Huy, et al. A complete survey on generative ai (aigc): Is chatgpt from gpt-4 to gpt-5 all you need? *arXiv preprint arXiv:2303.11717*, 2023.
- [295] D. J. Zhang, J. Z. Wu, J.-W. Liu, R. Zhao, L. Ran, Y. Gu, D. Gao, and M. Z. Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv:2309.15818*, 2023.
- [296] H. Zhang, Z. Wu, Z. Xing, J. Shao, and Y.-G. Jiang. Adadiff: Adaptive step selection for fast diffusion. *arXiv preprint arXiv:2311.14768*, 2023.
- [297] L. Zhang and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.
- [298] X. Zhang, Z. Wu, Z. Weng, H. Fu, J. Chen, Y.-G. Jiang, and L. S. Davis. Videolt: Large-scale long-tailed video recognition. In *ICCV*, 2021.
- [299] Y. Zhang, Y. Wei, D. Jiang, X. Zhang, W. Zuo, and Q. Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv:2305.13077*, 2023.
- [300] Z. Zhang, B. Li, X. Nie, C. Han, T. Guo, and L. Liu. Towards consistent video editing with text-to-image diffusion models. *arXiv preprint arXiv:2305.17431*, 2023.
- [301] Z. Zhang, B. Wu, X. Wang, Y. Luo, L. Zhang, Y. Zhao, P. Vajda, D. Metaxas, and L. Yu. Avid: Any-length video inpainting with diffusion model. In *CVPR*, 2024.
- [302] H. Zhao, K. Q. Lin, R. Yan, and Z. Li. Diffusionvmr: Diffusion model for video moment retrieval. *arXiv preprint arXiv:2308.15109*, 2023.
- [303] M. Zhao, R. Wang, F. Bao, C. Li, and J. Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv:2305.17098*, 2023.
- [304] Y. Zhao, E. Xie, L. Hong, Z. Li, and G. H. Lee. Make-a-protagonist: Generic video editing with an ensemble of experts. *arXiv:2305.08850*, 2023.
- [305] X. Zhong, Z. Li, S. Chen, K. Jiang, C. Chen, and M. Ye. Refined semantic enhancement towards frequency diffusion for video captioning. In *AAAI*, 2023.
- [306] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu, and J. Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv:2211.11018*, 2022.
- [307] L. Zhou, C. Xu, and J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018.
- [308] Q. Zhou, R. Li, S. Guo, Y. Liu, J. Guo, and Z. Xu. Cadm: Codec-aware diffusion modeling for neural-enhanced video streaming. *arXiv:2211.08428*, 2022.
- [309] Y. Zhou and N. Shimada. Vision+ language applications: A survey. In *CVPR*, 2023.
- [310] B. Zhu, F. Wang, T. Lu, P. Liu, J. Su, J. Liu, Y. Zhang, Z. Wu, Y.-G. Jiang, and G.-J. Qi. Poseanimate: Zero-shot high fidelity pose controllable character animation. *arXiv preprint arXiv:2404.13680*, 2024.
- [311] J. Zhu, H. Yang, H. He, W. Wang, Z. Tuo, W.-H. Cheng, L. Gao, J. Song, and J. Fu. Moviefactory: Automatic movie creation from text using large generative models for language and images. *arXiv:2306.07257*, 2023.
- [312] W. Zielonka, T. Bolkart, and J. Thies. Towards metrical reconstruction of human faces. In *ECCV*, 2022.

Received 18 Nov 2023; revised 18 Jun 2024; accepted 22 Aug 2024