

A Survey on Video Diffusion Models

Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu,
Hang Xu, Zuxuan Wu and Yu-Gang Jiang

Abstract—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>.

Index Terms—Survey, Video Diffusion Model, Video Generation, Video Editing, AIGC

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 [1–7] have emerged as particularly successful, rapidly supplanting methods based on generative adversarial networks (GANs) [8–12] and auto-regressive Transformers [13–16] 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 [17–20], dense prediction [21–25], and diverse areas such as video synthesis [26–31] and 3D generation [32–34].

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 [26, 27, 29–31, 35, 36], video editing [37–41], and video understanding [42–45].

With rapid advancement of video diffusion models [27] and their demonstration of impressive results, the endeavor to track and compare recent research on this topic gains great importance.

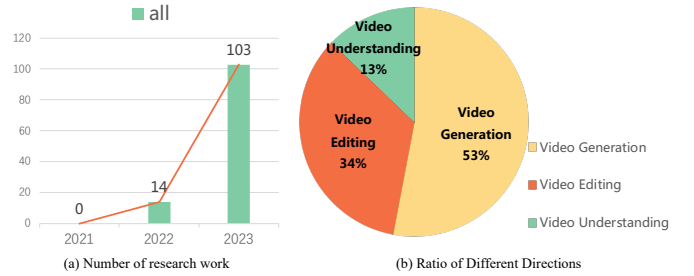


Fig. 1: Summarization on video diffusion model research works. (a) The number of related research works is rapidly increasing. (b) Video generation and editing are the top two research areas using diffusion models.

Several survey articles have covered foundational models in the era of AIGC [46, 47], encompassing the diffusion model itself [48, 49] and multi-modal learning [50–52]. There are also surveys specifically focusing on text-to-image [53] research and text-to-3D [54] applications. However, these surveys either provide only a coarse coverage of the video diffusion models or place greater emphasis on image models [49, 50, 53]. As such, in this work, we aim to fulfill the blank with a comprehensive review on 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,

- Z. Xing, Q. Feng, H. Chen, Z. Wu, Y-G. Jiang are with School of Computer Science, Fudan University. E-mail: {zxing20, zxwu, ygj}@fudan.edu.cn, {qjfeng21, chenhran21}@m.fudan.edu.cn
- Q. Dai, H. Hu are with Microsoft Research Asia. E-mail: {qid, hanhu}@microsoft.com
- H. Xu is with Huawei Noah's Ark Lab. E-mail: xbjxh@live.com

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.

2.1 Diffusion Model

Diffusion models [55, 56] 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) [57] in a variety of challenging tasks such as image generation [58–63], image super-resolution [60, 64–66], and image editing [67, 68]. Current research on diffusion models is mostly based on three predominant formulations: denoising diffusion probabilistic models (DDPMs) [55, 58, 69], score-based generative models (SGMs) [59, 61], and stochastic differential equations (Score SDEs) [56, 70].

2.1.1 Denoising Diffusion Probabilistic Models (DDPMs)

A *denoising diffusion probabilistic model* (DDPM) [55, 58, 69] 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) [59, 61] 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 $q(x_t|x_0) = \mathcal{N}(x_t; x_0, \sigma_t^2\mathbf{I})$, 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 [56] 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 analysis 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 [26, 28, 31], unconditional video generation [71–73], and text-guided video editing [37, 38, 74], *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 [75], animations [76, 77], virtual reality content, educational demonstration videos [78], *etc.*

- **Unconditional Video Generation** is a generative modeling task where the objective is to generate a continuous and visually coherent sequence of videos starting from random noise or a fixed initial state, without relying on specific input conditions. Unlike conditional video generation, unconditional video generation does not require any external guidance or prior information [27, 29, 79]. The generative model needs to autonomously 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 [29, 38, 80], 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 [81], small-scale datasets to specific domains [82, 83], resulting in relatively monotonous video generation. With the emergence of large-scale video-text paired datasets, tasks such as general text-to-video generation [26, 27] 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 [84–86], with small-scales

Dataset	Year	Text	Domain	#Clips	Resolution
MSR-VTT [84]	2016	Manual	Open	10K	240P
DideMo [85]	2017	Manual	Flickr	27K	-
LSMDC [86]	2017	Manual	Movie	118K	1080P
ActivityNet [87]	2017	Manual	Action	100K	-
YouCook2 [88]	2018	Manual	Cooking	14K	-
How2 [89]	2018	Manual	Instruct	80K	-
VATEX [90]	2019	Manual	Action	41K	240P
HowTo100M [91]	2019	ASR	Instruct	136M	240P
WTS70M [92]	2020	Metadata	Action	70M	-
YT-Temporal [93]	2021	ASR	Open	180M	-
WebVid10M [94]	2021	Alt-text	Open	10.7M	360P
Echo-Dynamic [95]	2021	Manual	Echocardiogram	10K	-
Tiktok [96]	2021	Manual	Action	0.3K	-
HD-VILA [97]	2022	ASR	Open	103M	720P
VideoCC3M [98]	2022	Transfer	Open	10.3M	-
HD-VG-130M [30]	2023	Generated	Open	130M	720P
InternVid [99]	2023	Generated	Open	234M	720P
CelebV-Text [100]	2023	Generated	Face	70K	480P

TABLE 1: The comparison of main caption-level video datasets.

Datasets	Year	Categories	#Clips	Resolution
UCF-101 [101]	2012	101	13K	256 × 256
Cityscapes [102]	2015	30	3K	256 × 256
Moving MNIST [81]	2016	10	10K	64 × 64
Kinetics-400 [103]	2017	400	260K	256 × 256
BAIR [104]	2017	2	45K	64 × 64
DAVIS [105]	2017	-	90	1280 × 720
Sky Time-Lapse [83]	2018	1	38K	256 × 256
Ssthv2 [106]	2018	174	220K	256 × 256
Kinetics-600 [107]	2018	600	495K	256 × 256
Tai-Chi-HD [82]	2019	1	3K	256 × 256
Bridge Data [108]	2021	10	7K	256 × 256
Mountain Bike [109]	2022	1	1K	576 × 1024
RDS [35]	2023	2	683K	512 × 1024

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

(less than 120K) and a limited focus on specific domains (*e.g.* movie [86], action [87, 92], cooking [88]). With the introduction of the open-domain WebVid-10M [94] 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 coverage of videos in the general text-to-video (T2V) tasks, VideoFactory [30] and InternVid [99] introduce larger-scale (130M & 234M) and high-definition (720P) open-domain datasets.

- **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. It is notable that several of these datasets are also applied to other tasks. For instance, UCF-101 [101], Kinetics [103, 107], and Something-Something [106] are typical benchmarks for action recognition. DAVIS [105] was initially proposed for the video object segmentation task and later became a commonly used benchmark for video editing. Among these datasets, UCF-101 [101] 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 [110] that encompasses 101 action categories, including human sports, musical instrument playing, and interactive actions. Akin to UCF, Kinetics-400 [103] and Kinetics-600 [107] are two datasets encompassing more complex action categories and

larger data scale, while retaining the same application scope as UCF-101 [101]. The Something-Something [106] dataset, on the other hand, possesses both category-level and caption-level labels, rendering it particularly suitable for text-conditional video prediction tasks [111]. 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 [82, 83], single-domain attributes (digital number [81], driving scenery [35, 102, 109], robot [108]) 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) [35] or extended duration [109, 112]. For example, Long Video GAN [109] proposes horseback dataset which has 66 videos with an average duration of 6504 frames at 30fps. Video LDM [35] collects RDS dataset consists of 683,060 real driving videos of 8 seconds 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 [26, 30, 31, 37], 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 [113]. 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) [114], Peak Signal-to-Noise Ratio (PSNR) [115], Structural Similarity Index (SSIM) [116], and CLIPSIM [117]. FID [114] 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 [118] to extract features from real and synthesized videos, and computing mean and covariance matrices. These statistics are then combined to calculate the FID [114] score.

Both SSIM [116] and PSNR [115] are pixel-level metrics. SSIM [116] evaluates brightness, contrast, and structural features of original and generated images, while PSNR [115] is a coefficient representing the ratio between peak signal and Mean Squared Error (MSE) [119]. 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 [117] is a method for measuring image-text relevance. Based on the CLIP [117] 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 [26, 30, 31, 35, 37, 120].

- **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) [121] is a video quality evaluation metric based on FID [114]. Unlike image-level methods that use the Inception [118] network to extract features from single frame, FVD [121] employs the Inflated-3D Convnets (I3D) [122] pre-trained on Kinetics [103] to extract features from video clips. Subsequently, FVD scores are computed through the combination of means and covariance matrices. Similar to FVD [121], Kernel Video Distance (KVD) [123] is also based on I3D [122] features, but it differentiates itself by utilizing Maximum Mean Discrepancy (MMD) [124], a kernel-based method, to assess the quality of generated videos. Video IS (Inception Score) [125] calculates the Inception score of generated videos using features extracted by the 3D-Convnets (C3D) [126], which is often applied in evaluation on UCF-101 [101]. 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 [117] is commonly used in video editing tasks [31, 37, 127] to measure the coherence of edited videos. Its calculation involves computing CLIP image embeddings for all frames of the edited videos and reporting the average cosine similarity between all pairs of video frames.

3 VIDEO GENERATION

In this section, we categorize video generation into four groups and provide detailed reviews for each: General 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 is demonstrated in Fig. 2.

3.1 Video Generation with Text Condition

Evidenced by recent research [1, 2, 171], the interaction between generative AI and natural language is of paramount importance. While significant progress has been achieved in generating images from text [1–3, 16], the development of Text-to-Video (T2V) approaches is still in its early stages. In this context, we first provide a brief overview of some non-diffusion methods [172, 173], followed by delving into the introduction of T2V models on both training-based and training-free diffusion techniques.

3.1.1 Non-diffusion T2V methods

Before the advent of diffusion-based models, early efforts in the field were primarily rooted in GANs [8], VQ-VAE [174] and auto-regressive Transformer [173] frameworks.

Among these works, GODIVA [175] is a representation work to use VQ-VAE [174] for general T2V task. It pretrains the model on Howto100M [91] that contains more than 100M video-text pairs. The proposed model shown excellent zero-shot performance at the time. Soon afterwards, auto-regressive Transformer methods lead the main-stream T2V task due to their explicit density modeling and stable training advantages compared with GANs [8, 10–12]. Among them, CogVideo [173] represents an extensive open-source video generation model that innovatively leverages the pretrained CogView2 [16] as its backbone for video generation tasks. In addition to the above stated works, PHENAKI [176] introduces a novel C-ViViT [177] backbone for variable length

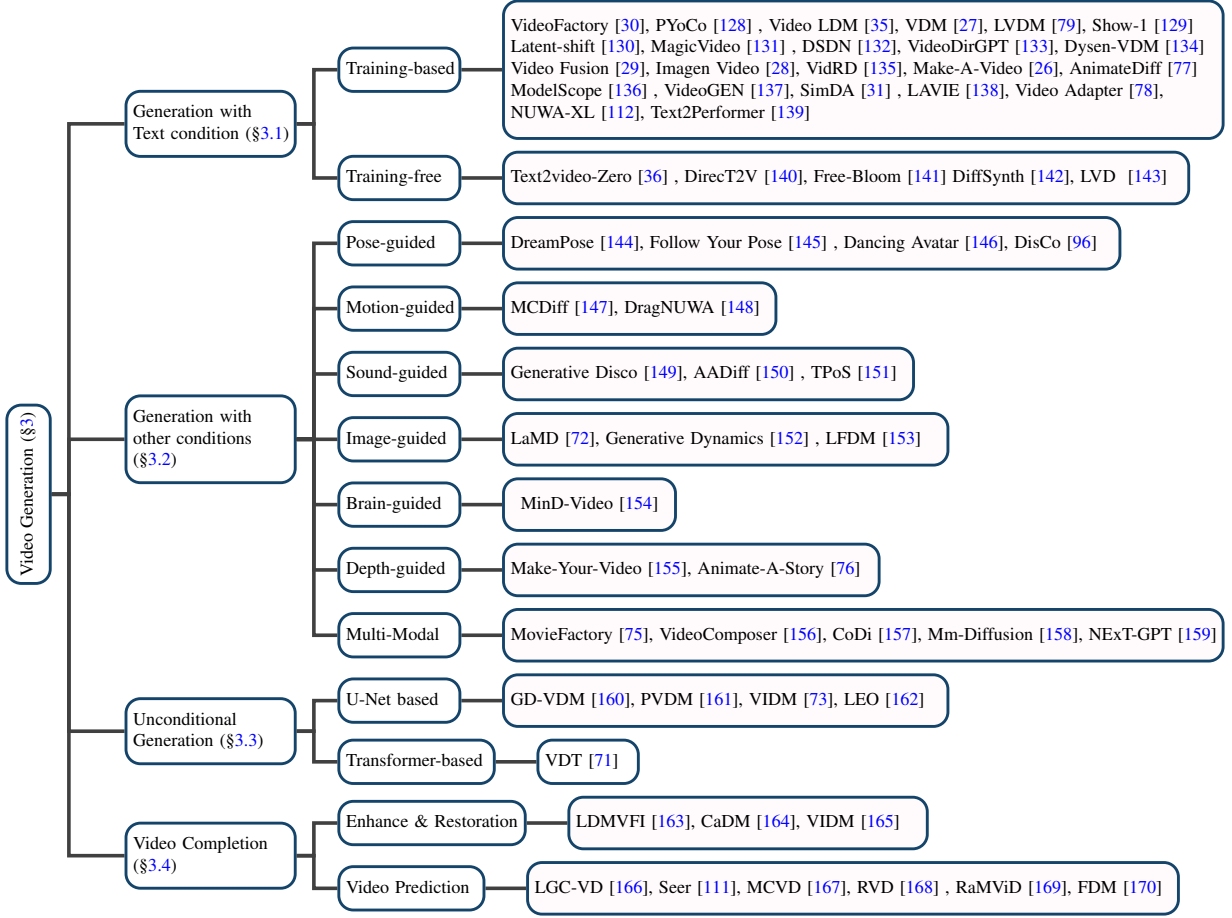


Fig. 2: Taxonomy of Video Generation. Key aspects of Video Generation include General T2V Generation, Domain-specific Generation, Conditional Control Generation, and Video Completion.

video generation. NUWA [172] is an unified model for T2I, T2V and video prediction tasks based on auto-regressive Transformer. MMVG [178] proposes an efficient mask strategy for several video generation tasks (T2V, video prediction and video refilling).

3.1.2 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 task.

• **Early T2V Exploration** Among the multitude of endeavors, VDM [27] stands as the pioneer in devising a video diffusion model for video generation. It extends the conventional image diffusion U-Net [179] 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 [27], which requires paired *video-text* datasets, Make-A-Video [26] 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 in pixel-level, MagicVideo [131] stands as one of the earliest works to employ the Latent Diffusion Model (LDM) [1] 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 [1] 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 [6] and Unconditional Guidance [180] 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.

ModelScope [136] incorporates spatial-temporal convolution and attention into LDM [1] for T2V tasks. It adopts a mixed training approach using LAION [181] and WebVid [94], and serves as an open-source baseline¹ method.

1. <https://modelscope.cn/models/damo/text-to-video-synthesis/summary>

Previous methods predominantly rely on 1D convolutions or temporal attention [131] to establish temporal relationships. Latent-Shift [130], on the other hand, focuses on lightweight temporal modeling. Drawing inspiration from TSM [182], it shifts channels between adjacent frames in convolution blocks for temporal modeling. Additionally, the model maintains the original T2I [1] capability while generating videos.

- **Multi-stage T2V methods** Imagen Video [28] extends the mature T2I model, Imagen [7], to the task of video generation. The cascaded video diffusion model is composed of seven sub-models, with one dedicated to base video generation, three for spatial super-resolution, and three for temporal super-resolution. Together, these sub-models form a comprehensive three-stage training pipeline. It validates the effectiveness of numerous training techniques employed in T2I training, such as classifier-free guidance [180], conditioning augmentation [6], and v-parameterization [183]. Additionally, the authors leverage progressive distillation techniques [183, 184] to speed up the sampling time of the video diffusion model. The multi-stage training techniques introduced therein have become effective strategies for mainstream high-definition video generation.

Concurrently, Video LDM [35] trains a T2V network composed with 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 [138] 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 [129] first introduces the fusion of pixel-based [185] and latent-based [1] diffusion models for T2V generation. Its framework comprises four distinct stages, with the initial three operating at a low resolution pixel-level: key frame generation, frame interpolation, and super resolution. Notably, pixel-level stages can generate videos with precise text alignment. The fourth stage is composed of a latent super-resolution module, which 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 [29] 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 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 like DALLÉ-2 [2] for training T2V models accelerates convergence, but its text embedding might face challenges in understanding long temporal sequences of text.

PYCo [128] 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 [186] 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.

- **Datasets Contribution** VideoFactory [30] takes note of the low resolution and watermark presence in the previously widely used WebVid [94] dataset. As a response, it constructs a large-scale video dataset, HD-VG-130M, consisting of 130 million video-text pairs from open-domain sources. This dataset is collected from HD-VILA [97] via BLIP-2 [187] caption, which claims high resolution and is devoid of watermarks. Additionally, VideoFactory introduces a swapped cross-attention mechanism to facilitate interaction between the temporal and spatial modules, resulting in improved temporal relationship modeling. Trained on this high-definition dataset, the approach presented in the paper is capable of generating high-resolution videos at (1376×768) resolution.

VidRD [135] introduces the Reuse and Diffuse framework, which iteratively generates additional frames by reusing the original latent representations and following the previous diffusion process. Furthermore, it utilizes static images, long videos and short videos when constructing the video-text dataset. For static images, dynamic aspects are introduced through random zoom or pan operations. Short videos [103] are annotated using BLIP-2 [187] labeling for categorization, while long videos [188] are first segmented and then annotated based on MiniGPT-4 [189] to retain the required video clips. The construction of diverse categories and distributions within video-text datasets proves to be effective of enhancing the quality of video generation.

- **Efficient Training** ED-T2V [190] utilizes LDM [1] 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 [31] devises a parameter-efficient training approach for T2V tasks by maintaining the parameter of T2I model [1] 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.

- **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 [77] notices the success of LoRA [191] and Dreambooth [5] 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.

- **Removing Artifacts** To address the issue of flickers and artifacts in T2V-generated videos, DSDN [132] 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 [137] first utilizes a T2I model [1] 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. Compared to previous methods, introducing a reference image improves visual fidelity and reduces artifacts, allowing the model to focus more on learning video dynamics.

- **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 [134] introduces a method that transforms textual information into dynamic scene graphs. Leveraging Large Language Model (LLM) [171], Dysen-VDM [134] identifies pivotal actions from 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 [133] 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 [192], 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 significant advantages in layout and motion control for complex dynamic video generation.

- **Domain-specific T2V Generation** Video-Adapter [78] 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 [193] and Bridge Data [108] scenarios.

NUWA-XL [112] employs a coarse-to-fine generative paradigm, facilitating parallel video generation. It initially employs global diffusion to generate keyframes, 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 [139] decomposes human-centric videos into appearance and motion representations. It first employs unsupervised training on natural human videos using a VQVAE [174] latent space to disentangle appearance and pose representations. Subsequently, it utilizes a continuous VQ-diffuser [194, 195] to sample continuous pose embeddings. Finally, the authors employ a motion-aware masking strategy in the spatio-temporal domain on the pose embeddings to enhance temporal correlations.

3.1.3 Training-free T2V Diffusion Methods

While former methods are all training-based T2V approaches that typically rely on extensive datasets like WebVid [94] or other video datasets [97, 99]. Some recent researches [36, 141] aim at reducing heavy training costs by developing training-free T2V approaches, as will be introduced next.

Text2Video-Zero [36] utilizes the pre-trained T2I model Stable Diffusion [1] 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 [4] and Instruct-Pix2Pix [17], enabling the controlled generation of videos.

DirectT2V [140] and Free-Bloom [141], on the other hand, introduce large language model (LLM) [192, 196] 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, DirectT2V [140] uses a novel value mapping and dual-softmax filtering approach. Free-Bloom [141] 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 [143] first utilizes LLM [192] 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 [142] proposes a latent in-iteration deflickering framework and a video deflickering algorithm to mitigate flickering and generate coherent videos. Moreover, it can be applied to various domains, including video stylization and 3D rendering.

3.2 Video Generation with other Conditions

Most of the previously introduced methods pertains 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. 3.

3.2.1 Pose-guided Video Generation

Follow Your Pose [145] 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 (Text-to-Image) model is finetuned using (image, pose) pairs, enabling pose-controlled generation. In the second stage, the model leverages unlabeled videos for learning temporal modeling by incorporating temporal attention and cross-frame attention mechanisms. This two-stage training imparts the model with both pose control and temporal modeling capabilities.

Dreampose [144] constructs a dual-path CLIP-VAE [117] image encoder and adapter module to replace the original CLIP text encoder in LDM [1] 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 [146] focuses on synthesizing human dance videos. It utilizes a T2I model [1] 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 [171] is utilized to enhance coherence between adjacent frames. Additionally, it leverages OpenPose ControlNet [4] to harness the ability to generate high-quality human body videos based on poses.

Disco [96] addresses a novel problem setting known as referring human dance generation. It leverages the ControlNet [4],

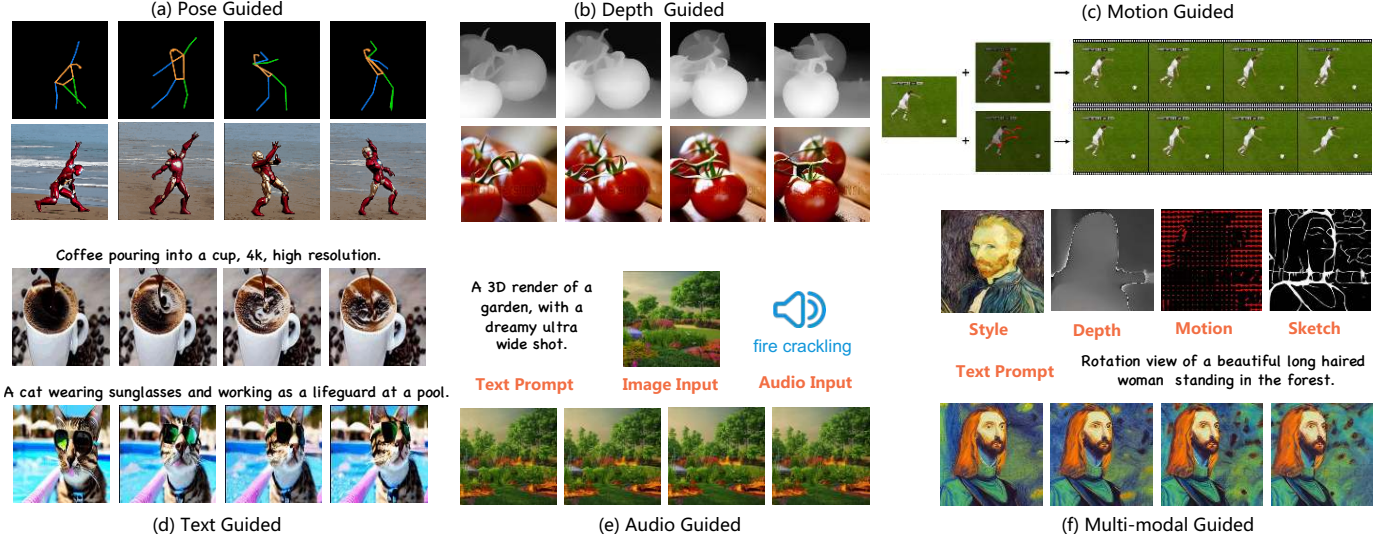


Fig. 3: Conditional video generation results with (a) Pose Guided [145], (b) Depth Guided [155], (c) Motion Guided [147], (d) Text Guided [31], (e) Audio Guided [151] and (f) Multi-modal Guided [156].

Grounded-SAM [197] and OpenPose [198] for background control, foreground extraction and pose skeleton extraction respectively. Moreover, large-scale image datasets [181, 199, 200] are employed for human attribute pre-training. By combining these training steps, Disco [96] lays a solid foundation for human-specific video generation tasks.

3.2.2 Motion-guided Video Generation

MCDiff [147] 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 [201] 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 [148] simultaneously introduce 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.

3.2.3 Sound-guided Video Generation

AADiff [150] 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 [202]. 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 [18] fashion to edit frames. This approach enables the generation of audio-synchronized videos without requiring any additional training.

Generative Disco [149] is an AI system designed for text-to-video generation aimed at music visualization. The system employs a pipeline that involves a large language model [192] followed by a text-to-image model [1] to achieve its goals.

TPoS [151] integrates audio inputs with variable temporal semantics and magnitude, building upon the foundation of the LDM [1] 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.

3.2.4 Image-guided Video Generation

LaMD [72] 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 [153] leverages conditional images and text for human-centric video generation. In the initial stage, a latent flow auto-encoder is trained to reconstruct videos. Moreover, a flow predictor [203] 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 [152] 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 [154] 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 [155] employs a novel approach for text-depth condition video generation. It integrates depth information as a conditioning factor by extracting it using MiDas [204] 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 [76], 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 [94]. Depth maps of these retrieved videos are obtained using offline depth estimation methods [204], 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 two-step approach enables the creation of personalized videos based on customized text descriptions.

3.2.7 Multi-modal guided Video Generation

VideoComposer [156] 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 [156] achieves higher video quality and improved detail in the generated content.

MM-Diffusion [158] represents the inaugural endeavor in 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 [158] also exhibits pronounced aptitude in effectuating video-to-audio translation.

MovieFactory [75] is dedicated to applying the diffusion model to the generation of film-style videos. It leverages ChatGPT [171, 205] to elaborate on user-provided text, creating comprehensive sequential scripts for the purpose of 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 [157] 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 [159] presents an end-to-end, any-to-any multi-modal LLM system. It integrates LLM [206] 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. During training, it fine-tunes only a small subset of parameters. Additionally, it introduces a modality-switching instruction tuning (MosIT) mechanism and manually curates a high-quality MosIT dataset. This dataset facilitates the acquisition of complex cross-modal semantic understanding and content generation capabilities.

3.3 Unconditional Video Generation

In this section, we delve into unconditional video generation. It refers to generating videos that belong to specific domain without extra condition. 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 later serves as a significant baseline method, VIDM [73] utilizes two streams: the content generation stream for video frame content generation, and the motion stream which defines video motion. By merging these two streams, consistent videos are generated. Furthermore, the authors employ Positional Group Normalization (PosGN) [69] to enhance video continuity and explore the combination of Implicit Motion Condition (IMC) and PosGN to address the generation consistency of long videos.

Similar to LDM [1], PVDM [161] 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.

Primarily focusing on synthesizing driving scene videos, GD-VDM [160] first generate 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 [162] 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** Different from most methods based on the U-Net [179] structure, VDT [71] pioneers the exploration of a video diffusion model grounded in the Transformer [207, 208] architecture. Leveraging the versatile scalability of Transformers, the authors investigate various temporal modeling approaches. Additionally, they apply VDT [71] to multiple tasks such as unconditional generation and video prediction.

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 [164] 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 [164] approach improves the compression efficacy of the encoder through the concurrent reduction of frame resolution and color bit-depth in video streams. Furthermore, CaDM [164] 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 [163] stands as the inaugural endeavor that employs a conditional latent diffusion model approach to address the video frame interpolation (VFI) task. In order to harness latent diffusion models for VFI, this work introduces a range of pioneering concepts. Notably, a video frame interpolation-specific autoencoding network is proposed, which integrates efficient self-attention modules and employs deformable kernel-based frame synthesis techniques to substantially enhance the performance.

VIDM [165] capitalizes on the pre-trained LDM [1] to address the task of video inpainting. By furnishing a mask for first-person perspective videos, the method leverages the image completion prior of LDM to generate inpainted videos.

3.4.2 Video Prediction

Seer [111] 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 [106, 108].

FDM [170] introduces a novel hierarchy sampling scheme for the purpose of long video prediction task. Additionally, a new CARLA [209] dataset is proposed. In comparison to auto-regressive methods, the proposed approach is not only more efficient but also yields superior generative outcomes.

MCVD [167] 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. The effectiveness of MCVD [167] is validated across various benchmarks [101, 102, 104] for both prediction and interpolation tasks.

Due to the tendency of autoregressive methods to yield implausible outcomes during the generation of lengthy videos, LGC-VD [166] 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 [168] (Residual Video Diffusion) adopts a diffusion model that utilizes the context vector of a convolutional Recurrent Neural Network (RNN) as condition 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. This work extensively compares with previous methods based on Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) across various benchmarks, providing substantial evidence of its efficacy.

RaMViD [169] 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 [104, 107], infilling [104], and upsampling [101, 107].

3.5 Benchmark Results

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

3.5.1 Zero-shot T2V Generation

- **Datasets.** General T2V methods, such as Make-A-Video [26] and VideoLDM [35], are primarily evaluated on the MSRVT [84] and UCF-101 [101] datasets in a zero-shot manner. MSRVT [84] 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 [101] 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 [121] and FID [114] metrics on the MSRVT [84] dataset. CLIPSIM [117] is used to measure the alignment between text and video. For the UCF-101 [101] dataset, the typical evaluation metrics include Inception Score [125], FVD [121], and FID [114] 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 MSRVT [84] and UCF-101 [101]. We also provide information about their parameter number, training data, extra dependencies, and resolution. It can be observed that methods relying on ChatGPT [134] or other input conditions [76, 155] exhibit a significant advantage over others, and the utilization of additional data [26, 30, 99] 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 focus on three specific datasets: UCF-101 [101], Taichi-HD [82], and Time-lapse [83]. These datasets

Method	Year	Training Data	Extra Dependency	Resolution	Params(B)	MSRVTT [84]			UCF-101 [101]		
						FID(↓)	FVD(↓)	CLIPSIM(↑)	FID(↓)	FVD(↓)	IS(↑)
Non-diffusion based method											
CogVideo [173]	2022	[94](5.4M)	-	256 × 256	15.5	23.59	1294	0.2631	179.00	701.59	25.27
MMVG [178]	2023	[94](2.5M)	-	256 × 256	-	-	-	0.2644	-	-	-
Diffusion based method											
LVDM [79]	2022	[94](2M)	-	256 × 256	1.16	-	742	0.2381	-	641.8	-
MagicVideo [131]	2022	[94](10M)	-	256 × 256	-	-	998	-	145.00	699.00	-
Make-A-Video [26]	2022	[94, 97]	-	256 × 256	9.72	13.17	-	0.3049	-	367.23	33.00
ED-T2V [190]	2023	[94](10M)	-	256 × 256	1.30	-	-	0.2763	-	-	-
InternVid [99]	2023	[94](10M) + 18M*	-	256 × 256	-	-	-	0.2951	60.25	616.51	21.04
Video-LDM [35]	2023	[94](10M)	-	256 × 256	4.20	-	-	0.2929	-	550.61	33.45
VideoComposer [156]	2023	[94](10M)	-	256 × 256	1.85	-	580	0.2932	-	-	-
Latent-shift [130]	2023	[94](10M)	-	256 × 256	1.53	15.23	-	0.2773	-	-	-
VideoFusion [29]	2023	[94](10M)	-	256 × 256	1.83	-	581	0.2795	75.77	639.90	17.49
Make-Your-Video [155]	2023	[94](10M)	Depth Input	256 × 256	-	-	-	-	-	330.49	-
PyoCo [128]	2023	[94] (22.5M)		256 × 256	-	9.73	-	-	-	355.19	47.76
CoDi [157]	2023	[94, 97]	-	512 × 512	-	-	-	0.2890	-	-	-
NExT-GPT [159]	2023	[94, 97]	-	320 × 576	1.83	13.04	-	0.3085	-	-	-
SimDA [31]	2023	[94](10M)	-	256 × 256	1.08	-	456	0.2945	-	-	-
Dysen-VDM [134]	2023	[94](10M)	ChatGPT	256 × 256	-	12.64	-	0.3204	-	325.42	35.57
VideoFactory [30]	2023	[94, 97]		-	256 × 256	2.04	-	-	0.3005	-	410.00
ModelScope [136]	2023	[94](10M)	-	256 × 256	1.70	11.09	550	0.2930	-	410.00	-
VideoGen [137]	2023	[94](10M)	Reference Image	256 × 256	-	-	-	0.3127	-	554.00	71.61
Animate-A-Story [76]	2023	[94](10M)		Depth Input	256 × 256	-	-	-	-	-	515.15
VidRD [135]	2023	[94, 103, 188](5.3M*)	-	256 × 256	-	-	-	-	-	363.19	39.37
LAVIE [138]	2023	[94](10M)+25M*	-	320 × 512	3.00	-	-	0.2949	-	526.30	-
VideoDirGPT [133]	2023	[94](10M)	GPT-4	256 × 256	1.92	12.22	550	0.2860	-	-	-
Show-1 [129]	2023	[94](10M)		320 × 576	-	13.08	538	0.3072	-	394.46	35.42
LVD [143]	2023	training-free		-	512 × 512	1.70	-	521	-	-	861.00

TABLE 3: Zero-shot Text-to-Video generation comparison on MSR-VTT [84] and UCF-101 [101] 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.

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 [102–104], 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 [101] dataset include IS [125] (Inception Score) and FVD [121] (Fréchet Video Distance). For the Time-lapse [83] and Taichi-HD [82] datasets, common evaluation metrics include FVD and KVD [123].

• **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, and extra dependencies are provided. It is evident that diffusion-based methods exhibit a significant advantage compared to traditional GANs [210, 212, 213] and autoregressive Transformer [173, 215] methods. Furthermore, if there is a large-scale pretraining or class conditioning, the performance tends to be further enhanced.

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 [74, 229, 232, 235], 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 is summarized in Fig. 4.

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. Yet, 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 [38] proposes a structure and content-aware model that provides full control over 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.

Method	Year	Type	Resolution	Extra	UCF-101 [101]		Taichi-HD [82]		Time-lapse [83]	
					FVD(↓)	IS(↑)	FVD(↓)	KVD(↓)	FVD (↓)	KVD(↓)
MoCoGAN [210]	2018	GAN	64 × 64	-	-	12.42	-	-	206.6	-
TGANv2 [125]	2020	GAN	128 × 128	-	-	26.60	-	-	-	-
StyleGAN-V [211]	2022	GAN	256 × 256	-	-	23.94	-	-	79.52	-
MoCoGAN-HD [212]	2021	GAN	256 × 256	-	700	33.95	144.7	25.4	183.6	13.9
DIGAN [213]	2022	GAN	128 × 128	-	577	32.70	128.1	20.6	114.6	6.8
StyleInV [214]	2023	GAN	256 × 256	-	-	-	186.72	-	77.04	-
MMVG [178]	2023	VQGAN	128 × 128	-	-	58.3	395	-	-	-
VideoGPT [215]	2021	Autoregressive	64 × 64	-	-	24.69	-	-	222.7	-
CCVS [216]	2021	Autoregressive	128 × 128	-	386	24.47	-	-	-	-
TATS [217]	2022	Autoregressive	128 × 128	Class Condition	278	79.28	94.6	9.8	132.6	5.7
CogVideo [173]	2022	Autoregressive	160 × 160	Pretrain+Class Condition	626	50.46	-	-	-	-
VDM [27]	2022	Diffusion	64 × 64	-	-	57.80	-	-	-	-
LVDM [79]	2022	Diffusion	256 × 256	-	372	27.00	99	15.3	95.2	3.9
VIDM [73]	2022	Diffusion	256 × 256	-	294.7	-	121.9	-	57.4	-
LEO [162]	2022	Diffusion	256 × 256	-	-	-	122.7	-20.49	57.4	-
VideoFusion [29]	2023	Diffusion	-128 × 128	-	220	72.22	56.4	6.9	47.0	5.3
PVDM [161]	2023	Diffusion	256 × 256	-	343.6	74.40	-	-	55.41	-
VDT [71]	2023	Diffusion	64 × 64	-	283.0	-	-	-	-	-
PYoCo [128]	2023	Diffusion	256 × 256	-	310	60.01	-	-	-	-
Dysen-VDM [134]	2023	Diffusion	256 × 256	ChatGPT	255.42	95.23	-	-	-	-
Latent-Shift [130]	2022	Diffusion	256 × 256	Class Condition	360	92.72	-	-	-	-
ED-T2V [190]	2023	Diffusion	256 × 256	Class Condition	320	83.36	-	-	-	-
Make-A-Video [26]	2023	Diffusion	256 × 256	Pretrain+Class Condition	81.25	-82.55	-	-	-	-
VideoGen [137]	2023	Diffusion	256 × 256	Pretrain+Class Condition	345	82.78	-	-	-	-

TABLE 4: Finetuned video generated results of UCF-101 [101], Taichi-HD [82] and Time-lapse [83]. We report the FVD, IS and KVD scores evaluation metric of clips with 16 frames. Besides, we also report the resolution of each video frame for each evaluation result.

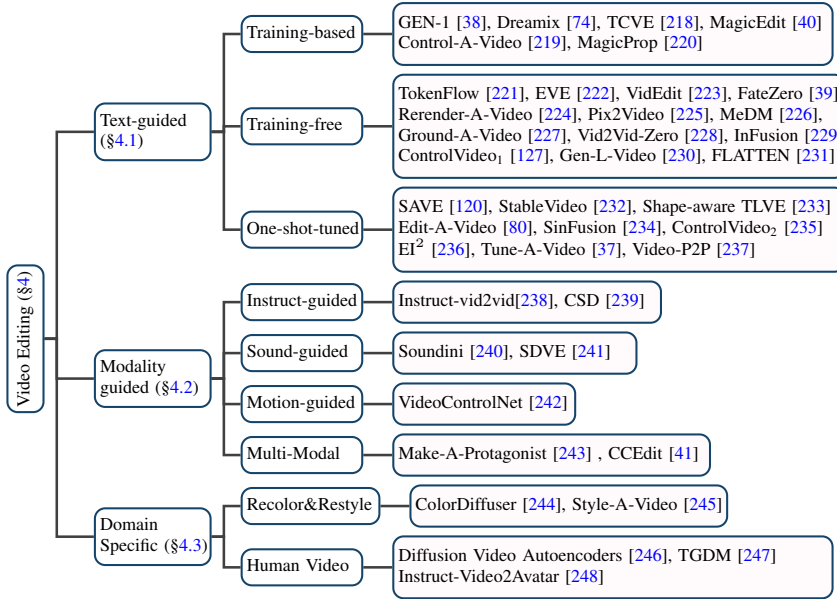


Fig. 4: Taxonomy of Video Editing. Key aspects of Video Editing include General Text-guided Video Editing, Modality-guided Video Editing and Domain-specific Video Editing.

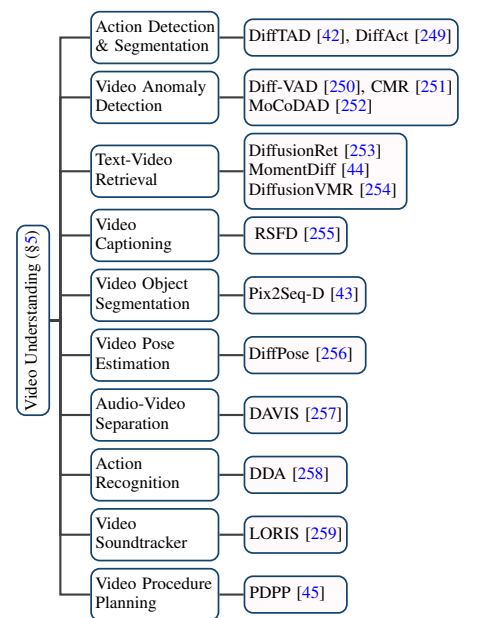


Fig. 5: Taxonomy of diffusion-based Video Understanding.

The high fidelity of Dreamix [74] results from two primary innovations: initializing generation using a low-resolution version of the original video and fine-tuning the generation model on the original video. They further propose a mixed fine-tuning approach with full temporal attention and temporal attention masking, significantly improving motion editability.

TCVE [218] proposes a Temporal U-Net, which effectively captures the temporal coherence of input videos. To connect the Temporal U-Net and the pre-trained T2I U-Net, the authors introduce a cohesive spatial-temporal modeling unit.

Control-A-Video [219] 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 (*i.e.*, generating video sequences based on the first frame), allowing Control-A-Video to produce videos of any length using an auto-regressive method.

Unlike most current methods simultaneously modeling appearance and temporal representation within a single framework, MagicEdit [40] innovatively separates the learning of content, structure, and motion for high fidelity and temporal coherence.

MagicProp [220] divides the video editing task into appearance editing and motion-aware appearance propagation, achieving temporal consistency and editing flexibility. They first select a frame from the input video and edit its appearance as a reference.

Then, they use an image diffusion model to auto-regressively generate the target frame, controlled by its previous frame, target depth, and reference appearance.

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 [221] demonstrates that consistency in edited videos can be achieved by enforcing consistency in the diffusion feature space. Specifically, this is accomplished by sampling key frames, jointly editing them, and propagating the features from the key frames 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 [223] combines atlas-based [260] and pre-trained T2I [1] 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 [224] employs hierarchical cross-frame constraints to enforce temporal consistency. The key idea involves using optical flow to apply dense cross-frame constraints, with the previously rendered frame serving as a low-level reference for the current frame and the first rendered frame acting as an anchor to maintain consistency in style, shape, texture, and color.

To address the issues of heavy costs in atlas learning [260] and per-video tuning [37], 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 [228] utilizes a null-text inversion [261] 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 [39], it also incorporates a spatial-temporal attention module.

Pix2Video [225] 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 [229] comprises two main components: first, it incorporates features from the residual block in decoder layers and attention features into the denoising pipeline for the editing prompt, highlighting its zero-shot editing capability. Second, it merges the attention for edited and unedited concepts by employing the mask extraction obtained from cross-attention maps, ensuring consistency.

ControlVideo₁ [127] directly adopts the architecture and weights from ControlNet [4], 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 key frames.

EVE [222] 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 [226] 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 [230] explores long video editing by treating long videos as temporally overlapping short videos. Through the proposed Temporal Co-Denoising methods, it extends off-the-shelf short video editing models [37, 79, 225] to handle editing videos comprising hundreds of frames while maintaining consistency.

To ensure consistency across all frames in the edited video, FLATTEN [231] incorporates optical flow into the attention mechanism of the diffusion model. The proposed Flow-guided attention allows patches from different frames to be placed on the same flow path within the attention module, enabling mutual attention and enhancing the consistency of video editing.

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 provides greater editing flexibility compared to training-free methods.

SinFusion [234] pioneers the one-shot-tuned diffusion-based models, which can learn the motions of a single input video from only a few frames. Its backbone is a fully convolutional DDPM [262] network, hence can be used to generate images of any size.

SAVE [120] 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 [80] 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 issue, they propose a masking method called *sparse-causal blending*, which automatically generates a mask to approximate the edited region.

Tune-A-Video [37] leverages a sparse spatio-temporal attention mechanism which 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 [237] 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 [235] mainly focuses on improving attention modules in the diffusion model and ControlNet [4]. 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 [233] utilizes the T2I model and handles shape changes by propagating the deformation field between the input and edited keyframe to all frames.

EI² [236] 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 [232] 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 generating video based on the given input video and instructions. Due to the lack of video-instruction datasets, InstructVid2Vid [238] leverages the combined use of ChatGPT, BLIP [187], and Tune-A-Video [37] 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 [239] 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 [17] to achieve coherent editing of multiple images with instruction.

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 [240] 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 [241] 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 [242] utilizes both diffusion model and ControlNet [4]. The method sets the first frame as the I-frame with the rest divided into different

group of pictures (GoP). The last frame of different GoPs is set as the P-frame while others are set as B-frames. Then, given an input video, the model first generates the I-frame directly based on the input’s I-frame using the diffusion model and ControlNet, followed by generating the P-frames through the motion-guided P-frame generation module (MgPG), in which the optical flow information is leveraged. Finally, the B-frames are interpolated based on the reference I/P-frame and the motion information instead of using the time-consuming diffusion model.

4.2.4 Multi-Modal Video Editing

Make-A-Protagonist [243] presents a multi-modal conditioned video editing framework to alter the protagonist. Specifically, they utilize BLIP-2 [187] for video captioning, CLIP Vision Model [117] and DALLE-2 Prior [2] for visual and textual clues encoding, and ControlNet [4] for the video consistency. During inference, they propose a mask-guided denoising sampling to combine experts to achieve without-annotation video editing.

CCEdit [41] decouples video structure and appearance for controllable and creative video editing. It preserves the video structure using the foundational ControlNet [4] while allowing appearance editing through text prompts, personalized model weights, and customized center frames. Additionally, the proposed temporal consistency modules and interpolation models can generate high-frame-rate videos seamlessly.

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 involves inferring plausible and temporally consistent colors for grayscale frames, which requires considering temporal, spatial and semantic consistency as well as color richness and faithfulness simultaneously. Built on the pre-trained T2I model, ColorDiffuser [244] proposes two novel techniques: the Color Propagation Attention as a replacement for optical flow, and Alternated Sampling Strategy to capture spatio-temporal relationships between adjacent frames.

- **Restyle** Style-A-Video [245] 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.

4.3.2 Human Video Editing

Diffusion Video Autoencoders [246] proposes a diffusion video autoencoder that extracts a single time-invariant feature (identity) and per-frame time-variant features (motion and background) from a given human-centric video, and further manipulates the single invariant feature for the desired attribute, which enables temporal-consistent editing and efficient computing.

In response to the increasing demand for creating high-quality 3D scenes easily, Instruct-Video2Avatar [248] takes in a talking head video and an editing instruction and outputs an edited version of 3D neural head avatar. They simultaneously leverage Instruct-Pix2Pix [17] for image editing, EbSynth [263] for video

stylization, and INSTA [264] for photo-realistic 3D neural head avatar.

TGDM [247] 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 fundamental video understanding tasks such as video temporal segmentation [42, 249], video anomaly detection [250, 251], text-video retrieval [44, 253], *etc.*, as will be introduced in this section. The taxonomy details of video understanding is summarized in Fig. 5.

5.1 Temporal Action Detection & Segmentation

Inspired by DiffusionDet [21], DiffTAD [42] 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 [265] architecture. Notably, the approach achieves state-of-the-art performance results on benchmarks such as ActivityNet [87] and THUMOS [266].

Similarly, DiffAct [249] 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 [267], 50Salads [268], and Breakfast [269].

5.2 Video Anomaly Detection

Dedicated to unsupervised video anomaly detection, DiffVAD [250] and CMR [251] harnesses 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 [270, 271] demonstrate the effectiveness of such paradigm, consequently significantly improving performance compared to prior research.

MoCoDAD [252] 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.

5.3 Text-Video Retrieval

DiffusionRet [253] formulates the retrieval task as a gradual process of generating a joint distribution $p(\text{candidates}, \text{query})$ from noise. During training, the generator is optimized using a generative loss, while the feature extractor is trained using a contrastive loss. In this manner, DiffusionRet [253] ingeniously combines the advantages of both generative and discriminative approaches and achieves outstanding performance in open domain scenarios, demonstrating its generalization ability.

MomentDiff [44] and DiffusionVMR [254] address the task of video moment retrieval, aiming to identify specific time intervals

in videos that correspond to given textual descriptions. Both approaches expand actual time intervals into random noise and learn to denoise the random noise back into the original time intervals. This process enables the model to learn a mapping from arbitrary random positions to actual locations, facilitating the accurate localization of video segments from random initialization.

5.4 Video Captioning

RSFD [255] 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.

5.5 Video Object Segmentation

Pix2Seq-D [43] redefines panoramic segmentation as a discrete data generation problem. It employs a diffusion model based on analog bits [272] to model panoptic masks, utilizing a versatile architecture and loss function. Furthermore, Pix2Seq-D [43] can model videos by incorporating predictions from previous frames, which enables the automatic learning of object instance tracking and video object segmentation.

5.6 Video Pose Estimation

DiffPose [256] addresses the problem of video-based human pose estimation by formulating it as a conditional heatmap generation task. Conditioned on the features generated in each denoising step, the method introduces a Spatio-Temporal representation learner that aggregates visual features across frames. Furthermore, a lookup-based multi-scale feature interaction mechanism is presented to create correlations across multiple scales for local joints and global contexts. This technique produces refined representations for keypoint regions.

5.7 Audio-Video Separation

DAVIS [257] 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. Due to its generative objective, DAVIS [257] is more appropriate for attaining high-quality sound separation across diverse categories.

5.8 Action Recognition

DDA [258] focuses on skeleton-based human action recognition. This method introduces diffusion-based data augmentation to obtain high-quality and diverse action sequences. It utilizes DDPMs [262] to generate synthesized action sequences, while the generation process is accurately guided by a spatial-temporal Transformer. Experimental results showcase the superiority of this approach in terms of naturalness and diversity metrics. Moreover, it confirms the effectiveness of applying synthesized high-quality data to existing action recognition models.

5.9 Video SoundTracker

LORIS [259] 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.

5.10 Video Procedure Planning

PDPP [45] focuses on procedure planning in instructional videos. The approach uses a diffusion model to depict the distribution of the entire intermediate action sequence, turning the planning problem into a sampling process from this distribution. Furthermore, accurate conditional guidance based on initial and final observations is provided using diffusion based U-Net model, enhancing the learning and sampling of action sequences from the learned distribution.

6 CHALLENGES AND FUTURE TRENDS

Despite the fact that diffusion-based methods have achieved significant advances in video generation, editing and understanding, there still exists certain open problems worthy of exploration. In this section, we summarize the current challenges and potential future directions.

- **Collecting Large-scale Video-Text Datasets** The substantial achievements in Text-to-Image synthesis are 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 [94] contains only 10 million instances and has a significant drawback of its limited visual quality, with a low resolution of 360P, further compounded by the presence of watermark artifacts. While efforts to explore new methods for obtaining datasets are in progress [30, 94, 135], there remains a pressing need for improvements in dataset scale, annotation accuracy, and video quality.

- **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 [35, 135]. Despite the efforts by methods such as SimDA [31] to mitigate training expenses, both the magnitude of dataset and temporal complexity remains a critical concern. Thus, exploring strategies for more efficient model training and reducing inference time is a valuable avenue for future research.

- **Benchmark and Evaluation Methods** Although benchmarks [84, 101] and evaluation methods [117, 121] for open-domain video generation exist, they are relatively limited in scope, as is demonstrated in [273]. 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) [121] and Inception Score (IS) [125] 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 a considerable reliance on user AB testing and subjective scoring, which is labor-intensive

and potentially biased due to subjectivity. Constructing more tailored evaluation benchmarks and metrics in the future is also a meaningful avenue of research.

- **Model Incapacity** While existing methods demonstrate remarkable progress, there are still numerous limitations due to model incapacity. For example, video editing methods often experience temporal consistency failures in certain cases, such as replacing human figures with animals. Additionally, we observe that for most methods discussed in Sec. 4.1, object replacement is limited to produce output of similar attributes. Moreover, in pursuing high fidelity, many current T2I-based models utilize key frames from the original video. However, due to the inherent limitations of off-the-shelf image generation models, injecting extra objects while preserving structural and temporal consistency remains unresolved. Further research and enhancement are essential to address these limitations.

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.

REFERENCES

- [1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *CVPR*, 2022. 1, 4, 5, 6, 7, 8, 9, 10, 13
- [2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” *arXiv:2204.06125*, 2022. 4, 6, 14
- [3] Midjourney. (2022) Midjourney. [Online]. Available: <https://www.midjourney.com/> 4
- [4] L. Zhang and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *ICCV*, 2023. 7, 13, 14
- [5] 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. 6
- [6] 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. 5, 6
- [7] 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. 1, 6
- [8] 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. 1, 4
- [9] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, “Alias-free generative adversarial networks,” in *NeurIPS*, 2021.
- [10] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *CVPR*, 2019. 4
- [11] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan,” in *CVPR*, 2020.
- [12] A. Sauer, K. Schwarz, and A. Geiger, “Stylegan-xl: Scaling stylegan to large diverse datasets,” in *ACM SIGGRAPH*, 2022. 1, 4

- [13] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *CVPR*, 2021. 1
- [14] 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.
- [15] 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.
- [16] M. Ding, W. Zheng, W. Hong, and J. Tang, "Cogview2: Faster and better text-to-image generation via hierarchical transformers," in *NeurIPS*, 2022. 1, 4
- [17] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *CVPR*, 2023. 1, 7, 14
- [18] 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. 8
- [19] 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.
- [20] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *CVPR*, 2023. 1
- [21] S. Chen, P. Sun, Y. Song, and P. Luo, "Diffusiondet: Diffusion model for object detection," in *ICCV*, 2022. 1, 15
- [22] Z. Gu, H. Chen, Z. Xu, J. Lan, C. Meng, and W. Wang, "Diffusioninst: Diffusion model for instance segmentation," *arXiv:2212.02773*, 2022.
- [23] J. Nam, G. Lee, S. Kim, H. Kim, H. Cho, S. Kim, and S. Kim, "Diff-match: Diffusion model for dense matching," *arXiv:2305.19094*, 2023.
- [24] 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.
- [25] 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. 1
- [26] 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. 1, 3, 4, 5, 10, 11, 12
- [27] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *NeurIPS*, 2022. 1, 3, 5, 12
- [28] 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. 3, 5, 6
- [29] 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. 1, 3, 5, 6, 11, 12
- [30] 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. 3, 4, 5, 6, 10, 11, 16
- [31] Z. Xing, Q. Dai, H. Hu, Z. Wu, and Y.-G. Jiang, "Simda: Simple diffusion adapter for efficient video generation," *arXiv:2308.09710*, 2023. 1, 3, 4, 5, 6, 8, 11, 16
- [32] 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. 1
- [33] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.
- [34] S. Luo and W. Hu, "Diffusion probabilistic models for 3d point cloud generation," in *CVPR*, 2021. 1
- [35] 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. 1, 3, 4, 5, 6, 10, 11, 16
- [36] 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. 1, 5, 7
- [37] 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. 1, 3, 4, 12, 13, 14
- [38] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," in *ICCV*, 2023. 3, 11, 12
- [39] 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. 12, 13
- [40] J. H. Liew, H. Yan, J. Zhang, Z. Xu, and J. Feng, "Magicedit: High-fidelity and temporally coherent video editing," *arXiv:2308.14749*, 2023. 12
- [41] R. Feng, W. Weng, Y. Wang, Y. Yuan, J. Bao, C. Luo, Z. Chen, and B. Guo, "Ccredit: Creative and controllable video editing via diffusion models," *arXiv:2309.16496*, 2023. 1, 12, 14
- [42] S. Nag, X. Zhu, J. Deng, Y.-Z. Song, and T. Xiang, "Diffad: Temporal action detection with proposal denoising diffusion," in *ICCV*, 2023. 1, 12, 15
- [43] 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. 12, 15
- [44] 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. 12, 15
- [45] H. Wang, Y. Wu, S. Guo, and L. Wang, "Pdpp: Projected diffusion for procedure planning in instructional videos," in *CVPR*, 2023. 1, 12, 16
- [46] 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. 1
- [47] J. Wu, W. Gan, Z. Chen, S. Wan, and H. Lin, "Ai-generated content (aigc): A survey," *arXiv:2304.06632*, 2023. 1
- [48] 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. 1
- [49] A. Ulhaq, N. Akhtar, and G. Pogrebna, "Efficient diffusion models for vision: A survey," *arXiv preprint arXiv:2210.09292*, 2022. 1
- [50] 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. 1
- [51] 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.
- [52] Y. Zhou and N. Shimada, "Vision+ language applications: A survey," in *CVPR*, 2023. 1
- [53] 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. 1
- [54] C. Li, C. Zhang, A. Waghware, 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. 1
- [55] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020. 2
- [56] 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. 2
- [57] 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. 2
- [58] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *ICML*, 2015. 2
- [59] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *NeurIPS*, 2019. 2
- [60] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022. 2
- [61] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in *NeurIPS*, 2020. 2
- [62] 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.
- [63] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," in *NeurIPS*, 2021. 2
- [64] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *TPAMI*, 2022. 2
- [65] G. Batzolis, J. Stanczuk, C.-B. Schönlieb, and C. Etmann, "Conditional image generation with score-based diffusion models," *arXiv:2111.13606*, 2021.
- [66] B. Kavar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *NeurIPS*, 2022. 2
- [67] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *CVPR*, 2022. 2
- [68] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "Ilvr: Conditioning method for denoising diffusion probabilistic models," in *ICCV*, 2021. 2
- [69] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *ICML*, 2021. 2, 9
- [70] Y. Song, C. Durkan, I. Murray, and S. Ermon, "Maximum likelihood training of score-based diffusion models," in *NeurIPS*, 2021. 2
- [71] 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. 3, 5, 9, 12

- [72] Y. Hu, Z. Chen, and C. Luo, “Lamd: Latent motion diffusion for video generation,” *arXiv:2304.11603*, 2023. 5, 8
- [73] K. Mei and V. Patel, “Vidm: Video implicit diffusion models,” in *AAAI*, 2023. 3, 5, 9, 12
- [74] 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. 3, 11, 12
- [75] 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. 3, 5, 9
- [76] 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. 3, 5, 9, 10, 11
- [77] 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. 3, 5, 6
- [78] 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. 3, 5, 7
- [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. 3, 5, 11, 12, 13
- [80] 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. 3, 12, 13
- [81] N. Srivastava, E. Mansimov, and R. Salakhudinov, “Unsupervised learning of video representations using lstms,” in *ICML*, 2015. 3, 4
- [82] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “First order motion model for image animation,” in *NeurIPS*, 2019. 3, 4, 10, 11, 12
- [83] 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. 3, 4, 10, 11, 12
- [84] J. Xu, T. Mei, T. Yao, and Y. Rui, “Msr-vtt: A large video description dataset for bridging video and language,” in *CVPR*, 2016. 3, 10, 11, 16
- [85] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, “Localizing moments in video with natural language,” in *ICCV*, 2017. 3
- [86] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, “Movie description,” *IJCV*, 2017. 3
- [87] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, “Dense-captioning events in videos,” in *ICCV*, 2017. 3, 15
- [88] L. Zhou, C. Xu, and J. Corso, “Towards automatic learning of procedures from web instructional videos,” in *AAAI*, 2018. 3
- [89] 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. 3
- [90] 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. 3
- [91] 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. 3, 4
- [92] 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. 3
- [93] 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. 3
- [94] 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. 3, 5, 6, 7, 9, 11, 16
- [95] 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. 3
- [96] 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. 3, 5, 7, 8
- [97] 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. 3, 6, 7, 11
- [98] 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. 3
- [99] 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. 3, 7, 10, 11
- [100] 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. 3
- [101] 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. 3, 4, 10, 11, 12, 16
- [102] 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. 3, 4, 10, 11
- [103] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *CVPR*, 2017. 3, 4, 6, 11
- [104] F. Ebert, C. Finn, A. X. Lee, and S. Levine, “Self-supervised visual planning with temporal skip connections,” *CoRL*, 2017. 3, 10, 11
- [105] 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. 3
- [106] 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. 3, 4, 10
- [107] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, “A short note about kinetics-600,” *arXiv:1808.01340*, 2018. 3, 10
- [108] 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. 3, 4, 7, 10
- [109] 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. 3, 4
- [110] YouTube. Youtube. [Online]. Available: <https://www.youtube.com/> 3
- [111] X. Gu, C. Wen, J. Song, and Y. Gao, “Seer: Language instructed video prediction with latent diffusion models,” *arXiv:2303.14897*, 2023. 4, 5, 10
- [112] 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. 4, 5, 7
- [113] H. Alqahtani, M. Kavakli-Thorne, G. Kumar, and F. SBSSTC, “An analysis of evaluation metrics of gans,” in *ICITA*, 2019. 4
- [114] G. Parmar, R. Zhang, and J.-Y. Zhu, “On aliased resizing and surprising subtleties in gan evaluation,” in *CVPR*, 2022, pp. 11 410–11 420. 4, 10
- [115] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *TIP*, 2004. 4
- [116] —, “Image quality assessment: from error visibility to structural similarity,” *TIP*, 2004. 4
- [117] 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. 4, 7, 10, 14, 16
- [118] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *CVPR*, 2016. 4
- [119] 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. 4
- [120] 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. 4, 12, 13
- [121] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, “Fvd: A new metric for video generation,” in *ICLR*, 2019. 4, 10, 11, 16
- [122] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *CVPR*, 2017. 4
- [123] 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. 4, 11
- [124] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” *J Mach Learn Res*, 2012. 4
- [125] M. Saito, S. Saito, M. Koyama, and S. Kobayashi, “Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan,” *IJCV*, 2020. 4, 10, 11, 12, 16
- [126] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *ICCV*, 2015. 4
- [127] 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. 4, 12, 13
- [128] 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. 5, 6, 11, 12
- [129] 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. 5, 6, 11
- [130] 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. 5, 6, 11, 12
- [131] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu, and J. Feng, "Magievideo: Efficient video generation with latent diffusion models," *arXiv:2211.11018*, 2022. 5, 6, 11
- [132] 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. 5, 6
- [133] H. Lin, A. Zala, J. Cho, and M. Bansal, "Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning," *arXiv:2309.15091*, 2023. 5, 7, 11
- [134] 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. 5, 7, 10, 11, 12
- [135] 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. 5, 6, 11, 16
- [136] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, "Modelscope text-to-video technical report," *arXiv:2308.06571*, 2023. 5, 11
- [137] 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:2309.00393*, 2023. 5, 7, 11, 12
- [138] 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. 5, 6, 11
- [139] Y. Jiang, S. Yang, T. L. Koh, W. Wu, C. C. Loy, and Z. Liu, "Text2performer: Text-driven human video generation," in *ICCV*, 2023. 5, 7
- [140] 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. 5, 7
- [141] H. Huang, Y. Feng, C. Shi, L. Xu, J. Yu, and S. Yang, "Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator," in *NeurIPS*, 2023. 5, 7
- [142] 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. 5, 7
- [143] L. Lian, B. Shi, A. Yala, T. Darrell, and B. Li, "Llm-grounded video diffusion models," *arXiv:2309.17444*, 2023. 5, 7, 11
- [144] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman, "Dreampose: Fashion image-to-video synthesis via stable diffusion," *arXiv:2304.06025*, 2023. 5, 7
- [145] 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. 5, 7, 8
- [146] 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. 5, 7
- [147] 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. 5, 8
- [148] 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. 5, 8
- [149] V. Liu, T. Long, N. Raw, and L. Chilton, "Generative disco: Text-to-video generation for music visualization," *arXiv:2304.08551*, 2023. 5, 8
- [150] S. Lee, C. Kong, D. Jeon, and N. Kwak, "Aadiff: Audio-aligned video synthesis with text-to-image diffusion," in *CVPRW*, 2023. 5, 8
- [151] 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. 5, 8
- [152] Z. Li, R. Tucker, N. Snively, and A. Holynski, "Generative image dynamics," *arXiv:2309.07906*, 2023. 5, 8
- [153] 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. 5, 8
- [154] Z. Chen, J. Qing, and J. H. Zhou, "Cinematic mindscapes: High-quality video reconstruction from brain activity," *arXiv:2305.11675*, 2023. 5, 8
- [155] J. Xing, M. Xia, Y. Liu, Y. Zhang, 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. 5, 8, 9, 10, 11
- [156] 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. 5, 8, 9, 11
- [157] Z. Tang, Z. Yang, C. Zhu, M. Zeng, and M. Bansal, "Any-to-any generation via composable diffusion," in *NeurIPS*, 2023. 5, 9, 11
- [158] 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. 5, 9
- [159] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "Next-gpt: Any-to-any multimodal llm," *arXiv:2309.05519*, 2023. 5, 9, 11
- [160] A. Lapid, I. Achituve, L. Bracha, and E. Fetaya, "Gd-vdm: Generated depth for better diffusion-based video generation," *arXiv:2306.11173*, 2023. 5, 9
- [161] S. Yu, K. Sohn, S. Kim, and J. Shin, "Video probabilistic diffusion models in projected latent space," in *CVPR*, 2023. 5, 9, 12
- [162] 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. 5, 9, 12
- [163] D. Danier, F. Zhang, and D. Bull, "Ldmvfi: Video frame interpolation with latent diffusion models," *arXiv preprint arXiv:2303.09508*, 2023. 5, 10
- [164] 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. 5, 10
- [165] M. Chang, A. Prakash, and S. Gupta, "Look ma, no hands! agent-environment factorization of egocentric videos," *arXiv:2305.16301*, 2023. 5, 10
- [166] S. Yang, L. Zhang, Y. Liu, Z. Jiang, and Y. He, "Video diffusion models with local-global context guidance," in *IJCAI*, 2023. 5, 10
- [167] V. Voleti, A. Jolicoeur-Martineau, and C. Pal, "Mcvd-masked conditional video diffusion for prediction, generation, and interpolation," in *NeurIPS*, 2022. 5, 10
- [168] R. Yang, P. Srivastava, and S. Mandt, "Diffusion probabilistic modeling for video generation," *arXiv:2203.09481*, 2022. 5, 10
- [169] T. Höppe, A. Mehrjou, S. Bauer, D. Nielsen, and A. Dittadi, "Diffusion models for video prediction and infilling," *Trans. Mach. Learn. Res.*, 2022. 5, 10
- [170] W. Harvey, S. Naderiparizi, V. Masrani, C. Weibach, and F. Wood, "Flexible diffusion modeling of long videos," in *NeurIPS*, 2022. 5, 10
- [171] OpenAI. (2022) Chatgpt: A large-scale generative model for conversational ai. [Online]. Available: <https://openai.com/chatgpt> 4, 7, 9
- [172] C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang, and N. Duan, "Nüwa: Visual synthesis pre-training for neural visual world creation," in *ECCV*, 2022. 4, 5
- [173] 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. 4, 11, 12
- [174] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *NeurIPS*, 2017. 4, 7
- [175] C. Wu, L. Huang, Q. Zhang, B. Li, L. Ji, F. Yang, G. Sapiro, and N. Duan, "Godiva: Generating open-domain videos from natural descriptions," *arXiv:2104.14806*, 2021. 4
- [176] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, and D. Erhan, "Phenaki: Variable length video generation from open domain textual description," in *ICLR*, 2023. 4
- [177] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *ICCV*, 2021. 4
- [178] 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. 5, 11, 12
- [179] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015. 5, 9
- [180] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS*, 2021. 5, 6
- [181] 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. 5, 8
- [182] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *ICCV*, 2019. 6
- [183] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *ICLR*, 2022. 6
- [184] C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, and T. Salimans, "On distillation of guided diffusion models," in *CVPR*, 2023. 6
- [185] D. Lab. (2023) Deepfloyd if. [Online]. Available: <https://github.com/deep-floyd/IF> 6
- [186] 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. 6
- [187] 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. 6, 14
- [188] 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. 6, 11
- [189] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “Minigt-4: Enhancing vision-language understanding with advanced large language models,” *arXiv:2304.10592*, 2023. 6
- [190] 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. 6, 11, 12
- [191] 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. 6
- [192] OpenAI, “Gpt-4 technical report,” *arXiv:2303.08774*, 2023. 7, 8
- [193] 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. 7
- [194] 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. 7
- [195] 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. 7
- [196] 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. 7
- [197] 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. 8
- [198] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in *CVPR*, 2017. 8
- [199] 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. 8
- [200] Y. Jafarian and H. S. Park, “Learning high fidelity depths of dressed humans by watching social media dance videos,” in *CVPR*, 2021. 8
- [201] 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. 8
- [202] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, “Clap learning audio concepts from natural language supervision,” in *ICASSP*, 2023. 8
- [203] 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. 8
- [204] 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. 9
- [205] 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. 9
- [206] 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. 9
- [207] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, “Transformer-xl: Attentive language models beyond a fixed-length context,” in *ACL*, 2019. 9
- [208] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *ICLR*, 2021. 9
- [209] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, “Carla: An open urban driving simulator,” in *CoRL*, 2017. 10
- [210] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, “Mocogan: Decomposing motion and content for video generation,” in *CVPR*, 2018. 11, 12
- [211] 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. 12
- [212] 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. 11, 12
- [213] 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. 11, 12
- [214] Y. Wang, L. Jiang, and C. C. Loy, “Styleinv: A temporal style modulated inversion network for unconditional video generation,” in *ICCV*, 2023. 12
- [215] W. Yan, Y. Zhang, P. Abbeel, and A. Srinivas, “Videogpt: Video generation using vq-vae and transformers,” *arXiv preprint arXiv:2104.10157*, 2021. 11, 12
- [216] G. Le Moing, J. Ponce, and C. Schmid, “Ccvs: context-aware controllable video synthesis,” in *NeurIPS*, 2021. 12
- [217] 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. 12
- [218] 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. 12
- [219] 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. 12
- [220] 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. 12
- [221] M. Geyer, O. Bar-Tal, S. Bagon, and T. Dekel, “Tokenflow: Consistent diffusion features for consistent video editing,” *arXiv:2307.10373*, 2023. 12, 13
- [222] 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. 12, 13
- [223] P. Couairon, C. Rambour, J.-E. Haugeard, and N. Thome, “Videdit: Zero-shot and spatially aware text-driven video editing,” *arXiv:2306.08707*, 2023. 12, 13
- [224] S. Yang, Y. Zhou, Z. Liu, and C. C. Loy, “Render a video: Zero-shot text-guided video-to-video translation,” in *SIGGRAPH Asia*, 2023. 12, 13
- [225] D. Ceylan, C.-H. P. Huang, and N. J. Mitra, “Pix2video: Video editing using image diffusion,” in *ICCV*, 2023. 12, 13
- [226] 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. 12, 13
- [227] H. Jeong and J. C. Ye, “Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models,” *arXiv:2310.01107*, 2023. 12
- [228] 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. 12, 13
- [229] A. Khandelwal, “Infusion: Inject and attention fusion for multi concept zero shot text based video editing,” in *ICCVW*, 2023. 11, 12, 13
- [230] 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. 12, 13
- [231] 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. 12, 13
- [232] W. Chai, X. Guo, G. Wang, and Y. Lu, “Stablevideo: Text-driven consistency-aware diffusion video editing,” in *ICCV*, 2023. 11, 12, 14
- [233] 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. 12, 14
- [234] Y. Ninkankin, N. Haim, and M. Irani, “Sinfusion: Training diffusion models on a single image or video,” in *ICML*, 2022. 12, 13
- [235] 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. 11, 12, 14
- [236] 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. 12, 14
- [237] S. Liu, Y. Zhang, W. Li, Z. Lin, and J. Jia, “Video-p2p: Video editing with cross-attention control,” *arXiv:2303.04761*, 2023. 12, 13
- [238] B. Qin, J. Li, S. Tang, T.-S. Chua, and Y. Zhuang, “Instructvid2vid: Controllable video editing with natural language instructions,” *arXiv:2305.12328*, 2023. 12, 14
- [239] S. Kim, K. Lee, J. S. Choi, J. Jeong, K. Sohn, and J. Shin, “Collaborative score distillation for consistent visual synthesis,” in *NeurIPS*, 2023. 12, 14
- [240] 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. 12, 14
- [241] 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. 12, 14

- [242] Z. Hu and D. Xu, "Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet," *arXiv:2307.14073*, 2023. [12](#), [14](#)
- [243] 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. [12](#), [14](#)
- [244] 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. [12](#), [14](#)
- [245] N. Huang, Y. Zhang, and W. Dong, "Style-a-video: Agile diffusion for arbitrary text-based video style transfer," *arXiv:2305.05464*, 2023. [12](#), [14](#)
- [246] 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. [12](#), [14](#)
- [247] 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. [12](#), [15](#)
- [248] S. Li, "Instruct-video2avatar: Video-to-avatar generation with instructions," *arXiv:2306.02903*, 2023. [12](#), [14](#)
- [249] D. Liu, Q. Li, A. Dinh, T. Jiang, M. Shah, and C. Xu, "Diffusion action segmentation," in *ICCV*, 2023. [12](#), [15](#)
- [250] A. O. Tur, N. Dall'Asen, C. Beyan, and E. Ricci, "Exploring diffusion models for unsupervised video anomaly detection," *IEEE VCIP*, 2023. [12](#), [15](#)
- [251] —, "Unsupervised video anomaly detection with diffusion models conditioned on compact motion representations," 2023. [12](#), [15](#)
- [252] 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. [12](#), [15](#)
- [253] 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. [12](#), [15](#)
- [254] H. Zhao, K. Q. Lin, R. Yan, and Z. Li, "Diffusionvmr: Diffusion model for video moment retrieval," *arXiv preprint arXiv:2308.15109*, 2023. [12](#), [15](#)
- [255] 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. [12](#), [15](#)
- [256] 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. [12](#), [15](#)
- [257] 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. [12](#), [15](#)
- [258] 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. [12](#), [15](#)
- [259] J. Yu, Y. Wang, X. Chen, X. Sun, and Y. Qiao, "Long-term rhythmic video soundtracker," in *ICML*, 2023. [12](#), [16](#)
- [260] O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel, "Text2live: Text-driven layered image and video editing," in *ECCV*, 2022. [13](#)
- [261] 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. [13](#)
- [262] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020. [13](#), [15](#)
- [263] O. Jamriska, "Ebsynth: Fast example-based image synthesis and style transfer," 2018. [14](#)
- [264] W. Zielonka, T. Bolkart, and J. Thies, "Towards metrical reconstruction of human faces," in *ECCV*, 2022. [15](#)
- [265] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020. [15](#)
- [266] 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. [15](#)
- [267] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *CVPR*, 2011. [15](#)
- [268] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *ACMSE*, 2013. [15](#)
- [269] H. Kuehne, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *CVPR*, 2014. [15](#)
- [270] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *CVPR*, 2018. [15](#)
- [271] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—a new baseline," in *CVPR*, 2018. [15](#)
- [272] T. Chen, R. Zhang, and G. Hinton, "Analog bits: Generating discrete data using diffusion models with self-conditioning," in *ICLR*, 2023. [15](#)
- [273] I. Chivileva, 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. [16](#)