

Sora as an AGI World Model? A Complete Survey on Text-to-Video Generation

JOSEPH CHO, Kyung Hee University, South Korea

FACHRINA DEWI PUSPITASARI, KAIST, South Korea

SHENG ZHENG, Kyung Hee University, South Korea

JINGYAO ZHENG, The Hong Kong Polytechnic University, Hong Kong SAR

LIK-HANG LEE, The Hong Kong Polytechnic University, Hong Kong SAR

TAE-HO KIM, Nota Inc., South Korea

CHOONG SEON HONG, Kyung Hee University, South Korea

CHAONING ZHANG*, Kyung Hee University, South Korea

The evolution of video generation from text, starting with animating MNIST numbers to simulating the physical world with Sora, has progressed at a breakneck speed over the past seven years. While often seen as a superficial expansion of the predecessor text-to-image generation model, text-to-video generation models are developed upon carefully engineered constituents. Here, we systematically discuss these elements consisting of but not limited to core building blocks (vision, language, and temporal) and supporting features from the perspective of their contributions to achieving a world model. We employ the PRISMA framework to curate 97 impactful research articles from renowned scientific databases primarily studying video synthesis using text conditions. Upon minute exploration of these manuscripts, we observe that text-to-video generation involves more intricate technologies beyond the plain extension of text-to-image generation. Our additional review into the shortcomings of Sora-generated videos pinpoints the call for more in-depth studies in various enabling aspects of video generation such as dataset, evaluation metric, efficient architecture, and human-controlled generation. Finally, we conclude that the study of the text-to-video generation may still be in its infancy, requiring contribution from the cross-discipline research community towards its advancement as the first step to realize artificial general intelligence (AGI).

CCS Concepts: • Computing methodologies → Computer vision tasks; Natural language generation; Machine learning approaches.

Additional Key Words and Phrases: Survey, Text-to-Video Generation, Text-to-Image Generation, Generative AI, Sora Model, Temporal Dynamics, Scalability in AI, Artificial General Intelligence, AI Models Generalization

1 INTRODUCTION

On February 15th, 2024, OpenAI introduced a new vision foundation model that can generate video from users' text prompts. The model named Sora, which people call a video version of ChatGPT, has raised excitement mainly from industries such as marketing [189], education [19], and filmmaking [182] as it promotes democratization of high-quality content creation that would normally require substantial resources. OpenAI claimed that Sora, due to being trained on a large-scale dataset of text-video pairs, has impressive near-real-world generation capability. This includes creating vivid characters, simulating smooth motion, depicting emotions, and provisioning detailed objects and backgrounds. Given these assertions, we are interested in exploring *how text-to-video generation models have come closer to being world models from a technical perspective*.

Cho and Puspitasari contribute equally. *Correspondence Author: Chaoning Zhang (chaoningzhang1990@gmail.com).

Authors' addresses: Joseph Cho, Kyung Hee University, South Korea, joyousaf@khu.ac.kr; Fachrina Dewi Puspitasari, KAIST, South Korea, puspitasaki-dewi@outlook.com; Sheng Zheng, Kyung Hee University, South Korea, zszhx2021@gmail.com; Jingyao Zheng, The Hong Kong Polytechnic University, Hong Kong SAR, jingyao.zheng@connect.polyu.hk; Lik-Hang Lee, The Hong Kong Polytechnic University, Hong Kong SAR, lik-hang.lee@polyu.edu.hk; Tae-Ho Kim, Nota Inc., South Korea, lik-hang.lee@polyu.edu.hk; Choong Seon Hong, Kyung Hee University, South Korea, cshong@khu.ac.kr; Chaoning Zhang*, Kyung Hee University, South Korea, chaoningzhang1990@gmail.com.

1.1 Brief Overview

Text-to-Video Generation Models. Emerging from text-to-image generation, text-to-video generation models expand the technological features of the image counterpart to handle the temporal dimension existing in video data. Similar to their text-to-image generation counterparts, text-to-video generation models also employ generative machine learning architectures such as VQ-VAE, GAN, autoregressive models, and diffusion models. To train the text-to-video generation model, pairs of text-video data are fed into the model, which triggers it to learn the approximation of the true data distribution and make inferences from unseen video. Here, the text prompt supplied by the user functions as the condition for generation to ensure that the output does not deviate from the intended classes implied by the prompt.

World Model. A generative artificial intelligence (AI) model that understands real-world mechanisms is often referred to as a world model. For the vision model, this apprehension of the world shall be reflected in various aspects of generation output, such as physics understanding, user visual comfort, the logic of content composition, etc. To achieve this capability, often the initial requirements that the model needs to pass are scalability and generalizability. Scalability refers to how much data is fed as input and whether the model shows a sign of emergent capability not observed in ordinary generation models. Generalizability refers to the ability of such a model to generate output beyond the training data distribution.

1.2 Method

We conduct a comprehensive review of studies in text-to-video generation models to thoroughly discuss their enabling technologies. Our survey utilizes Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [146] framework. We primarily collect conference and journal papers from well-recognized databases, including IEEE Xplorer, ACM Library, Scopus, and arXiv. The venues of the publications include but are not limited to AAAI, ACL, CVPR, ECCV, ICCV, ICLR, IJCAI, NAACL, NeurIPS, ACM Multimedia, and IEEE Transactions on Multimedia. We conducted a search on arXiv and the above databases on March 18th, 2024. Note that we include arXiv in our search library list since research in deep learning, particularly the computer vision domain, has developed faster than the peer-reviewed venues can provide. Given that text-to-video synthesis is still a growing study in the research community, we did not restrict the range of publication years of the papers collected. Using the search keyword of “text-to-video”, we initially curated 197 articles after briefly reviewing the fitness of the publication title. To ensure that we only review studies closely related to our survey objective, we devise several exclusion criteria, as follows:

- articles that discuss video synthesizing not conditioned on text prompt,
- articles that discuss text-video relationship other than the generation task (e.g, retrieval, editing, captioning),
- survey and review articles, and
- article from arXiv that has not received its first citation despite already being published more than a year ago (to proxy the evaluation towards the quality and usefulness of the papers).

Implementing this list in the selection process of the content of the abstract and full article, we finally obtained a final list of 97 papers used as the survey’s main articles. Figure 1 plots the statistics of these articles. This implies that text-to-video generation models have developed rapidly since 2023. Further, it indicates that the majority of these works were published as pre-prints in arXiv, which supports our decision to include arXiv as our search library despite its identity as a non-peer-reviewed publishing platform.

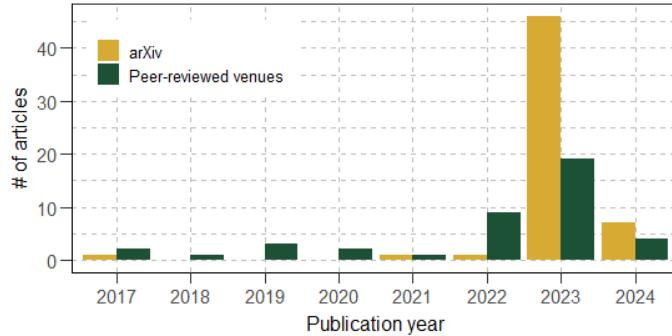


Fig. 1. Statistics of main articles curated to write this survey.

1.3 Contribution

According to the aforementioned statistics, text-to-video generation is still a relatively novel exploration in the research community. Consequently, there are only a handful of survey or review articles for this field. Table 1 lists a summary of these surveys in comparison with ours. These prior surveys reviewed works in text-to-video generation models using the approach of either summarizing key models or centering the discussion on a certain foundation model. Different from them, we provide a more comprehensive review of text-to-video generation technology, emphasizing not only the vision part but also other substantial parts such as language, temporal, and supporting features. Our work also complements the existing survey in the generative AI models for text-to-text [225], text-to-image [226], text-to-3D [111], and text-to-speech [227]. To summarise, our survey contributes the following to the research community in text-to-video generation models.

- We provide exhaustive technical discussion underpinning text-to-video generation models from multiple aspects, including vision processors, language interpreters, temporal handlers, other supporting features, and datasets and metrics commonly used.
- We anchor this discussion on the core topic of world modeling through the lens of the text-to-video generation task, which become increasingly important in today's generative AI landscape.
- We conduct thorough observation on the advancements and limitations of today's text-to-video generation models and advocate both potential applications and future research direction.

1.4 Structure

To present a comprehensive survey on the text-to-video generation model, we start by briefly introducing its underlying primary building blocks consisting of language interpreters, vision processors, and temporal handlers (§ 2). Further, we summarize other auxiliary functions implemented by these models to bring the output video closer to the definition of real-world illustration (§ 3). We also explore various datasets used for training and evaluating the text-to-video generation models, as well as metrics commonly employed to measure the model's performance (§ 4). We next present various potential applications of text-to-video generation models and their implications for world models as well as their ethical and social impacts (§ 5). Finally, our discussion section suggests interesting future research directions that society may exercise to circumvent challenges that still hinder the realization of world modeling through the text-to-video generation task (§ 6).

Table 1. Comparison of the extent of discussion between our survey and existing review papers.

Article	Summary of discussion	Technological discussion				
		Vis.	Lang.	Temp.	Feat.	D & M
[139]	Comparison of GAN-based text-to-video generation	GAN	•	-	-	•
[172]	Comparison of major text-to-video generation models	•	-	-	-	-
[236]	Combination of generative AI and LLM for video technologies and applications	•	LLM	-	-	-
[125]	Review of enabling technologies underlying Sora	diffusion	-	•	•	-
[177]	Survey of text-to-video generation models from Sora perspective	•	-	-	•	•
[94]	Comparison of major text-to-video generation foundation models	diffusion	-	-	-	-
Ours	Survey of technological foundations of text-to-video generation models and their world modeling roles	•	•	•	•	•

*Vis., Lang., Temp., Feat., and D & M refer to Vision processor, Language interpreter, Temporal handler, Supporting features, and Datasets and Metrics respectively.

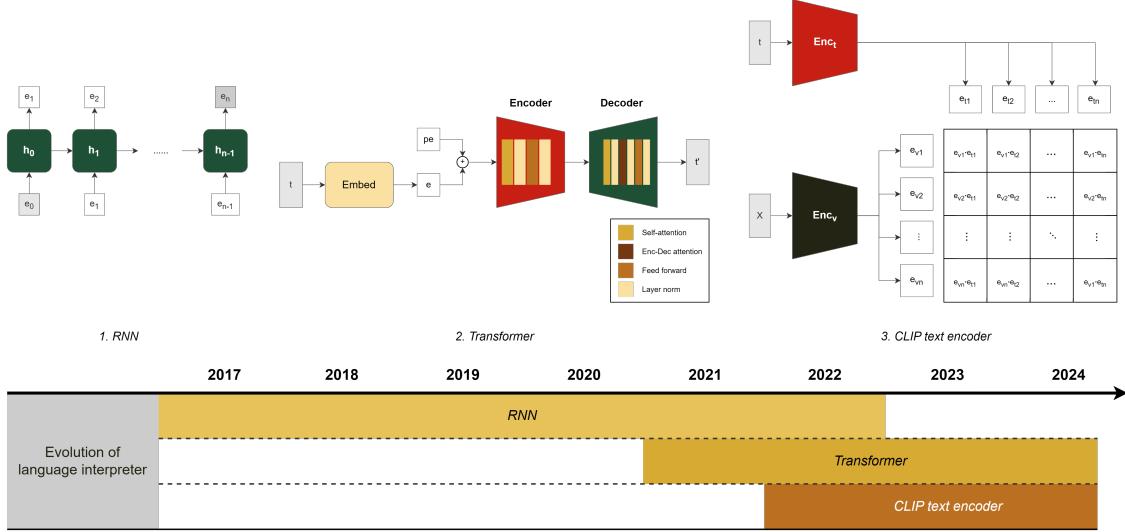
** [•] and [-] indicate available and unavailable discussions on the survey papers, respectively.

2 PRIMARY BUILDING BLOCKS

As its name echoes, text-to-video generation models involve three primary technological constructs, which are vision processors, language interpreters, and temporal handlers. Here, we discuss the common technological backbone used in text-to-video generation models. For each backbone, we briefly discuss the underlying concept of its mechanism and mention the text-to-video generation model that pioneered the implementation of such a backbone as well as other follower models.

2.1 Language Interpreter

Generating from text prompts requires the model to integrate a language interpretation model. These models (Figure 2) translate words into visual objects and coherently connect the text context with the nuance depicted in the image and the dynamic presented in the video.



* X , t , e , and pe refer to pixel space data, token, embedding, and position encoding respectively.

Fig. 2. Evolution of language interpreter used to process textual input in text-to-video generation models.

2.1.1 Recurrent Model. Recurrent networks are mainly used as the text prompt encoder by text-to-video generation models that employ GAN architecture. These models utilize simple RNN [10, 115, 116], LSTM [124, 143, 147], or GRU [97, 109] to encode input text prompt. As one of the earliest attempts at generating video from text, GAN-based models follow the practice of manually encoding the sequential scene generation from a general topic sentence, for which recurrent language models are needed. Text prompt is first converted into text embedding using vectorizers like GloVe [151], Skip-thought vector [99], CNN, or MLP. The resulting embedding is then sequentially encoded by the recurrent networks to extract the contextual understanding contained in the text.

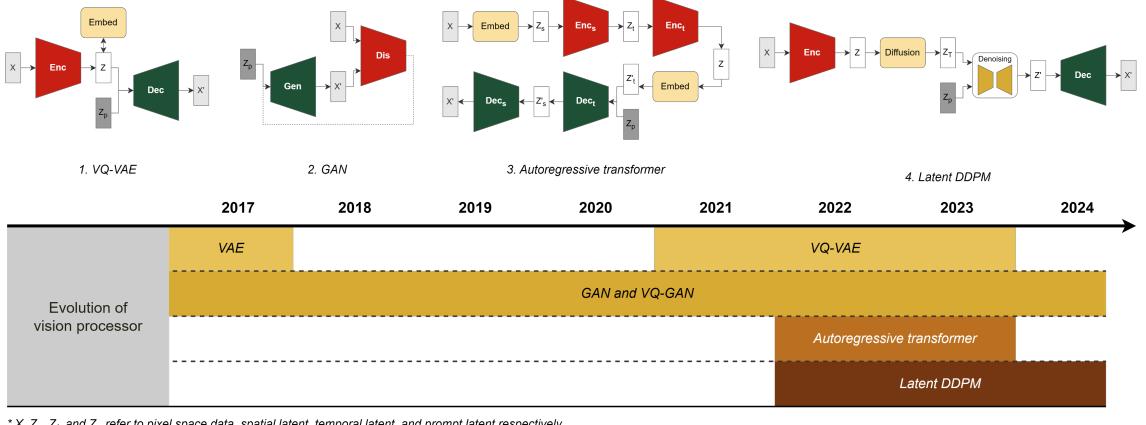
2.1.2 Transformer Model. The slightly modern architecture of text-to-video generation models, such as those that are based on autoregressive and vector-quantized (and a few diffusion) models, utilize a transformer [186] model to turn text prompts into language tokens. BERT and T5 are the two most common language encoders being integrated into these generation models. BERT [41] is an encoder-only transformer model that performs bidirectional attention with a random masking feature to the input token sequence that allows each token to attend both to the preceding and following token. Meanwhile, T5 [158] is an encoder-decoder transformer model designed for text-to-text transfer whose mechanism is akin to machine translation [8]. Although both BERT and T5 are conceptually similar in encoding text through denoising or masked token prediction objectives, T5 inherently has a larger parameter than BERT due to its architecture. Moreover, extensive experiments on T5 have shown that scaling up substantially increases the model performance [158]. For this reason, T5-family models are preferred by powerful text-to-video generation models such as Phenaki [188] and others with autoregressive [68, 88, 101] or diffusion [55, 56, 216, 229] architecture.

2.1.3 Contrastive Model. While the transformer models are excellent in performing sequence recognition from text description and correlating that to fine-grained details in the video being generated, these models are limited in the global understanding of the whole context in the video [11]. This might be one of the reasons a large number of text-to-video generation models (the rest of the papers that we reviewed) utilize a CLIP text encoder to decode the text description into the video. CLIP [157] is originally a vision-language representation model that was trained using pairs of image and caption data using contrastive learning. Meanwhile, the CLIP text encoder is essentially a transformer model. As CLIP is primarily devised as a zero-shot classifier, the text encoder was pretrained to match the class of a text description with the image as a whole. Despite the detailed-general representation trade-off, this training objective allows the CLIP to match image and text efficiently [157]. Such a characteristic might explain why the CLIP text encoder is utilized in many vision generation tasks, including text-to-video generation.

2.2 Vision Processor

Text-to-video generation is a computer vision task, thus the model’s main mechanism lies in its ability to comprehend the visual elements that exist in the video. As video is basically a sequence of images, many of these generation models implement similar vision processing models as commonly used in image generation models. Figure 3 illustrates the architecture and the evolution timeline of these processors.

2.2.1 VQ-VAE. Many text-to-video generation models perform the generation process in the latent space instead of the original pixel space to allow for less costly training. For this, VQ-VAE [185] is often used to encode the video into the latent variables \mathbf{Z} . VQ-VAE works with variational inference principle, similar to VAE [98]. The fundamental principle of variational inference is to get new data \mathbf{X}' through the reconstruction of \mathbf{Z} following approximation of posterior distribution $P(\mathbf{Z}|\mathbf{X})$ by $Q(\mathbf{Z})$ which can be conditioned to follow Gaussian distribution using regularization term.



* X , Z_s , Z_t , and Z_p refer to pixel space data, spatial latent, temporal latent, and prompt latent respectively.

Fig. 3. Evolution of vision backbone used to process visual input in text-to-video generation models.

For vision data, \mathbf{Z} is generally a continuous representation. Nevertheless, VQ-VAE compresses \mathbf{X} into \mathbf{Z} in a discrete space through nearest neighbor look-up at the embedding space learned using vector quantization. Discretization allows the model to be adaptable to other modalities (e.g., language and speech) and to fit real-world problems (e.g., reasoning and prediction) [185]. Sync-draw [143] is the first text-to-video model developed on VAE. Meanwhile, for VQ-VAE, GODIVA [205] pioneered the development which was later followed by several text-to-video generation models [3, 85, 86, 92, 136, 214]. They integrate VQ-VAE to encode the input video into discrete video tokens, which are then concatenated with the text embedding to be processed for video generation.

2.2.2 GAN. GAN is primarily utilized in text-to-video generation models to produce video whose frames have both high visual quality and diversity. Diverse new data \mathbf{X}' generation in GAN is possible as it is influenced by the objective of the generator θ_g is to maximize the likelihood of the discriminator being wrong in distinguishing between real and fake samples, $D(G(\mathbf{Z})) \approx 1$. Such a training objective coupled with regularization-free generation encourages GAN to attend more to the fine-grained quality of \mathbf{X}' , which makes the image produced have high visual quality. For these capabilities, GAN is used mostly in story visualization tasks as it helps the model generate diverse scenes based on the story flow. This task is pioneered by StoryGAN [115] which inspired other models to follow [109, 135, 180]. Not only in story visualization but GAN is also used in text-to-video generation tasks, benefiting mainly from its high generation quality. TGANS-C [147] and T2V [116] initiated this idea which is then followed by many other models that manipulate moving pictures [10, 97, 124, 138] and simulate body parts motion [100, 102, 173, 230]. Despite its high visual quality output, GAN falls short in its ability to generate high-resolution images. As the pixel size grows, generation with vanilla GAN is computationally infeasible because of these two reasons, *first*, GAN performs generation in the highly costly pixel space, and *second*, CNN [105] backbone is less expressive in learning the relational composition among visual elements [47]. To overcome this challenge, a few recent text-to-video generation models [51, 199, 233], led by MMVG [74], shifted to using VQ-GAN. VQ-GAN [47] is slightly different from VQ-VAE in that it extends the idea of quantization to discrete latent space by performing auto-regressive codebook learning using a transformer [186]. Thus, using the quantized latent space as an input, adversarial training with GAN further promotes the utilization

of perceptual loss, which encourages the latent synthesizing activity to learn richer codebook representations. This enables the transformer to probe further into the contextual understanding of elements in an image.

2.2.3 Autoregressive Transformer. Synthesizing in discrete latent space has been proven effective in joint learning of text and video. A few recent text-to-video generation models implement the straightforward idea of synthesizing video using only a transformer. Phenaki [188] was the first to propose this idea through its C-ViT architecture that modifies ViViT by adding causal attention, allowing auto-regressivity in the time dimension. ViViT [6] working principle is different from ViT [44]. ViViT fuses both spatial and temporal information during tokenization instead of tokenizing only spatial information and fusing temporal information later during concatenation. Using a transformer to handle video in this way has been shown to be better than the diffusion model because discretization has many benefits, such as supporting multiple modes of communication, speeding up (de)compression, and helping people understand context [220]. These reasons encourage the development of more transformer-based text-to-video generation models [68, 88, 101].

2.2.4 Diffusion. Since most text-to-video generation models use the diffusion model, it has recently become a prima donna. Majority of works that we curated (see Table A1 in Appendix A) are built on the foundation of Stable Diffusion [162]. Stable diffusion is a diffusion model that performs generation from input data \mathbf{x} in continuous latent space after being encoded by a VAE encoder. As a generative model, diffusion model or DDPM [81] has been proven to outperform GAN [42] as it offers a better balance between diversity and fidelity. DDPM synthesizes a new sample from a Gaussian noise that is generated through a forward diffusion process $q(\mathbf{z}_t | \mathbf{z}_0)$ that converts input data \mathbf{z}_0 into noise $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ by gradually adding an infinitesimal amount of noise for a T timesteps. The key generation process is done through a reverse denoising process $p_\theta(\mathbf{z}_{t-1} | \mathbf{x}_t)$ that iteratively generates less noisy data \mathbf{z}_{t-1} from \mathbf{z}_t using a neural network $\epsilon_\theta(\mathbf{z}_t, t)$. This neural network is optimized through the following objective function:

$$L_t^{simple} = \mathbb{E}_{t \sim [1, T], \mathbf{x}_0, \epsilon_t} \left[\| \epsilon_t - \epsilon_\theta(\sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \epsilon_t, t) \|^2 \right]$$

Modern DDPM allows users to control the generation process by inserting prompts such as text, audio, and localization marks (e.g., bounding box, segmentation mask, depth map). The classifier-free diffusion guidance handles all of these extra inputs. It is trained along with the DDPM neural network during the reverse denoising process [82].

2.3 Temporal Handler

As video is a sequence of images, a temporal handler is a critical element that complements the vision processor. While the latter focuses on learning the visual content within each frame, the former learns the dynamics of these contents following frame progression. Such a temporal model is unique to text-to-video generation models and can be performed with a range of mechanisms. Figure 4 illustrates common temporal handlers used in text-to-video generation models.

2.3.1 Temporal Attention. Adding a temporal layer is the most straightforward way of incorporating temporal dimensionality into the existing text-to-vision generation task. This approach is mostly exercised by text-to-video generation models whose architecture is naturally autoregressive (i.e., autoregressive and VQ-VAE). The temporal layer can be explicitly integrated into the generative transformer through several applications, including the temporal dimension of an axial transformer [86, 101, 205], spatiotemporal attention [68, 85], and causal transformer in the encoder module [88, 188]. There are also more subtle ways to apply temporal attention to text-to-video generation models. One way is to use neural ODEs [27] that approximate temporal dynamics [214], or another way is to use a bidirectional masked attention transformer that patchifies the input frames, thus turning them into temporal sequence [3, 92]. Due to its

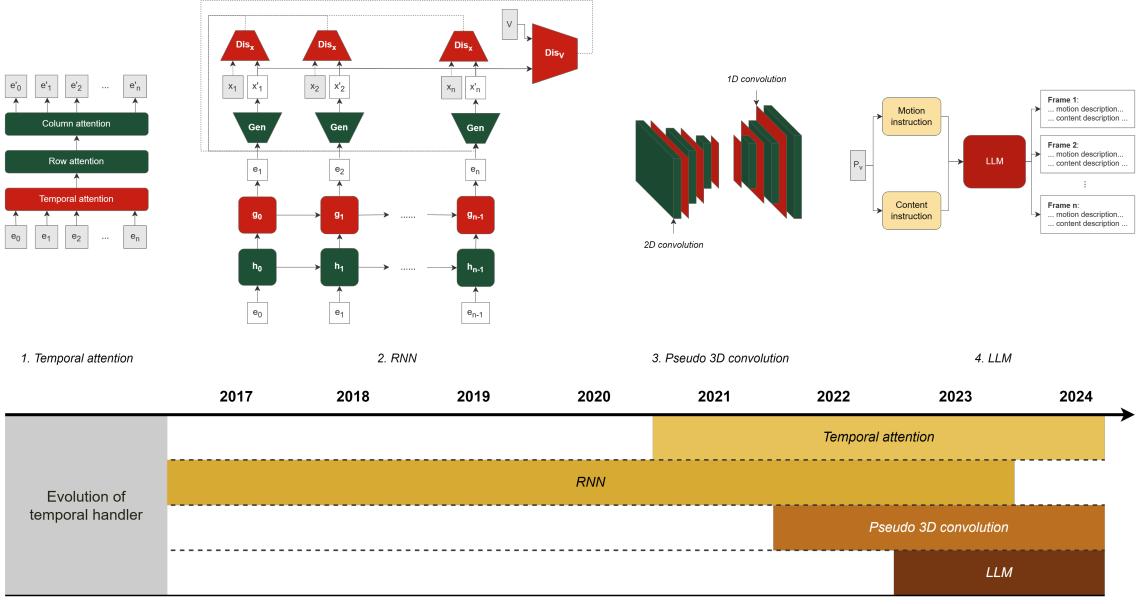


Fig. 4. Evolution of temporal handler used to align frames in text-to-video generation models.

forthright mechanism, the temporal handling method using temporal attention ensures the consistency yet diversity of the generated frames simply by relying on the tokenization of the input frames during training and injection of additional conditions such as context memory and motion anchor.

2.3.2 Recurrent Neural Network. For text-to-video generation models that do not have autoregressive architecture, such as GAN models, attaching a recurrent neural network is a common solution to handle temporal dimension. LSTM [84] and GRU [35] are the two most customary RNNs for generating temporal sequencing from the input text prompt. The RNN takes as input encoded text representation t_f , noise $\mathbf{z}_f \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and its previous hidden state h_{f-1} to produce the hidden state h_f which later is sent to the generator to output the video frame at time f . To ensure both temporal consistency and content diversity, GAN-based text-to-video generation models often incorporate an additional module that serves as a hint generator that connects the generated content across frames. Such a module may be devised in many approaches, such as through additional frame discriminator [97, 147], gist layer that blends and dynamically updates local and global context information [115, 116, 180], and MART [108] that memorizes the content described in the global narrative while decomposing it into frame-level captions [135, 180].

2.3.3 Pseudo-3D Convolutions and Attention. Network inflation is a familiar technique for incorporating temporal dimensionality into the diffusion model. Instead of directly inflating a 2D U-Net into a 3D U-Net, the network undergoes a pseudo-inflation, where a 1D convolution (temporal) layer is attached after every 2D convolution (spatial) layer. Through the pseudo-3D convolution layer, the input video of shape $batch \times frames \times channels \times height \times width$ is processed by the spatial layer as $(b \times f) \times c \times h \times w$ and the temporal layer as $b \times c \times f \times h \times w$. This separable convolution [36] technique not only greatly reduces the computational burden from direct replacement by a 3D convolution layer, but also preserves the knowledge of the pre-trained 2D generation model while simultaneously

updating the temporal parameter from scratch. Additionally, the temporal attention layer shares the knowledge that the temporal layer has acquired with all of the input elements. Similar to the pseudo-3D convolutions, the 1D (temporal) attention layer is also attached after the 2D (spatial) attention layer, making the whole attention block a pseudo-3D attention layer. This pseudo-3D attention layer technique inspired by VDM [83] also takes sinusoidal positional embedding to attach the frame indices information to the input tensor. Pioneered by Make-A-Video [171], these two network expansion methods have seemingly become a standard (80% of the papers we reviewed) adaptation technique to DDPM to accommodate video generation. Moreover, to ensure both temporal consistency and generation diversity inter-frames, DDPM models often couple the network expansion technique with various consistency modeling such as noise scheduling [55, 132, 156, 210, 216, 229], alignment modules [4, 154, 211], decoupled learning [26, 122, 155, 191], trajectory anchoring [24, 31, 96, 113, 118, 218], and temporal expansion of decoder module [15, 59].

2.3.4 Large Language Model. Borrowing the extensive performance of LLM in multimodal and multitask learning, very recent text-to-video generation models [87, 128, 130, 154] borrow LLM’s capability to encode simple text instruction into comprehensive scene descriptions. As straightforward as it sounds, the sequence of scene descriptions produced by the LLM is directly fed into the generation module, which primarily operates on text-to-image generation tasks. Meanwhile, a few other models [49, 117, 127] employ a more subtle approach of incorporating LLM as the temporal encoder. This approach utilizes LLM to generate scene information that is fed to the generation module as an auxiliary condition aside from the main text prompt and step size, much like injecting motion anchor to the generation module.

3 AUXILIARY TECHNIQUES

The attempt to generate video from text has existed since 2017, or even further back. Early days were dominated by simple tasks such as visualizing stories or moving still images. Story visualization attempts to illustrate a sequence of scenes in a short story using visualization. Although akin to animation, story visualization has less to do with optimizing the video frame rate due to its nature of generation. Meanwhile, moving still images, although simple, require the generated video to maintain a certain frame rate to visualize object motion smoothly. Studies in animating images have touched upon various subtasks such as changing object position [143, 147], animating human lip movement or facial expression [102, 230], or simulating simple human body movement [10, 97]. To date, the development of text-to-video generation models has integrated various functions to produce video content that is closest possible to human convenience comprehension. Among these are the injection of temporal conditions, utilization of efficient learning techniques, and evaluation of generated content with a feedback loop.

3.1 Frame Sequencing

Text is not the sole condition that can be supplied into the model to guide the video generation process. Several high- and low-level signals are exercisable for anchoring users’ intents in shaping the final generated content. Particularly in text-to-video generation tasks, conditions for temporal progression are considered the most prominent ones. We illustrate common temporal conditions integrated in the input of text-to-video generation model in Figure 5.

3.1.1 Trajectory Anchoring. Anchoring trajectory is considered the most apparent method in providing inter-frame temporal guidance for text-to-video models. This practice is commonly done by embedding physical signals in the pixel space representing the foreground subjects who perform particular motions. For instance, the spatial location of the subject is located by a bounding box [91, 133] that gradually changes position in each frame. Besides, other common

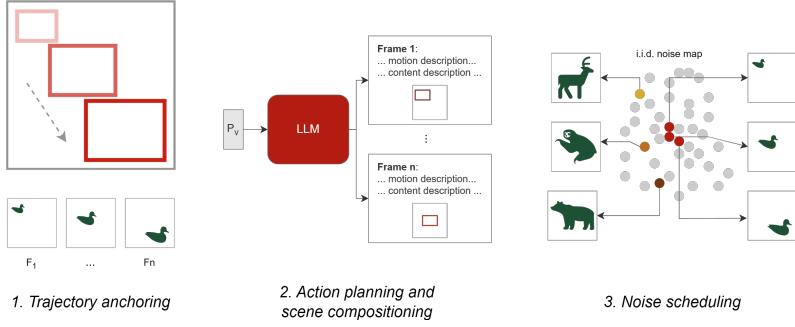


Fig. 5. Methods to inject temporal conditions into text-to-video generation models.

spatial signals such as segmentation mask [96], depth map [30, 50], optical flow [118], and scribble [43] can also be utilized.

3.1.2 Action Planning and Scene Composition. Visualizing temporal changes may implicitly be guided through a thorough textual description. The challenge of providing text-to-video generation models with comprehensive descriptions of the scene composition lies in the complexity of the text prompt that the user needs to devise. One way to solve this is to obtain the assistance of LLM. Given its superior ability in complex textual understanding, LLM can turn a simple user caption into a lengthy yet comprehensive text comprising scene description, entity categories and localizations, background scenery, and scene consistency grouping [49, 119, 128, 237]. Oftentimes, prompt expansion by LLM can be generated in multiple sequences where each sequence expands the preceding one [87, 127]. In a few cases, LLM can assist either in writing code that is used to operate a physics engine simulator such as Blender [130] or in sketching the spatial signal that indicates the object’s temporal propagation [117].

3.1.3 Noise Scheduling. In diffusion models, noise is an integral component of input that determines the content generated. As a rule of thumb, the diffusion model expects the input noise to come from Gaussian distribution to enable the generation of diverse output. Nevertheless, in text-to-video generation models, naively selecting i.i.d Gaussian noises may result in a sub-optimal performance as video requires inter-frame consistency [87]. Noises initialized for all frames interact with each other through a temporal attention mechanism [15]. Indeed, initializing noise for an individual frame is crucial for determining its content appearance. Nevertheless, noises from all frames combined must either be spatially clustered (high cosine similarity) [55] or obey a certain ordering to ensure temporal coherence [156]. Given this finding, noise injection has recently become a part of fundamental research avenues in text-to-video generation models. Scheduling noises for video generation can be done in several ways. Progressive scheduling is the earliest approach introduced by PYoCo [55]. The method generates noise ϵ for the subsequent frame in an autoregressive manner in which the ϵ_t is generated by perturbing ϵ_{t-1} [132]. Subsequently, different studies proposed different approaches to noise scheduling including calibrating noise at the terminal step to be zero [56], shuffling noises from a subset of frames [156], shifting noise following motion flow [30, 128], and joint noise sampling for all frames [87].

3.2 Efficient Learning

The major substantial challenges in performing text-to-video generation learning include data scarcity and an arduous iterative process. Thus, recent advances in model development have sought ways to circumvent this challenge through

performing multimodal joint training, attaching an adapter to an existing model, devising an efficient denoising process, and decoupling learning of different aspects (Figure 6).

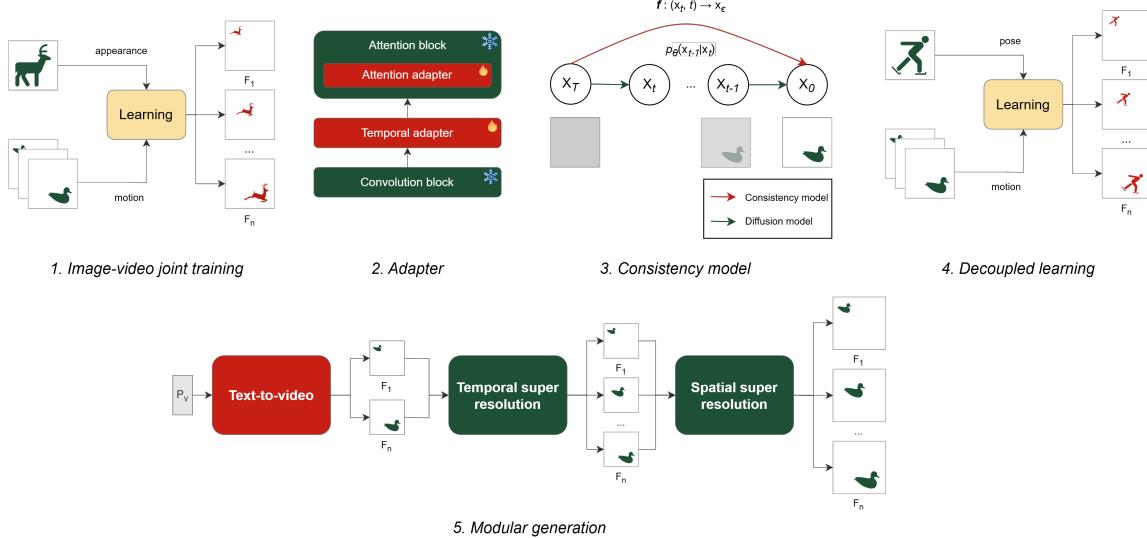


Fig. 6. Methods to efficiently train text-to-video generation models.

3.2.1 Image-Video Joint Training. As text-to-video generation is fundamentally developed from text-to-image generation, the number of text-image pair datasets is substantially larger than that of text-video datasets. While curating a large amount of text-video pair data may be the simplest solution, the coverage of the concept it can represent is considerably smaller than what can be represented by the text-image counterpart. Thus, the common practice in bypassing this requirement is through joint training of both text-image and text-video datasets. The model learns rich visual semantics from the text-image pair dataset and mines comprehensive motion understanding from the text-video dataset. This practice was introduced by Phenaki [188], which assigns a certain proportion of image-video in one training process. This method is followed by other models [68, 88, 155, 192, 194, 200, 223] afterward. Besides training at one go, it is also possible to jointly train image-video datasets subsequently [25, 26, 228, 237].

3.2.2 Adapter. Often, text-to-video generation models are the extension of powerful text-to-image generation models. Reforming these models entirely to fit the video generation objective and training them from scratch may cause catastrophic forgetting and lead to sub-optimal performance. Given this reason, several text-to-video generation approaches opt to attach adapters that handle temporal dimension to the already powerful text-to-image generation models [211]. These adapters are often known as motion modules. Apart from inserting adapters that function as temporal handlers, other generation models may attach adapters whose objective is to close the distribution gap between the text-image dataset used in the pretrained text-to-image generation model and the text-video dataset used for inflating such a model to text-to-video generation one [65, 67, 211].

3.2.3 Consistency Model. The main requirement allowing the diffusion model to generate video in a Gaussian way is to remove the noise infinitesimally through numerous denoising steps. Nevertheless, this requirement makes such a

process computationally expensive, and it is even more exorbitant for processing video data. To alleviate such a costly essential, research in text-to-image generation models has recently proposed a consistency model [129]. Consistency model [174] is a method for the diffusion model used for directly mapping input noise to data. Here, the denoising process can be completed in a single step and only sampling is performed in multiple steps to ensure that output is of high quality despite efficient computation. In the video generation task, the integration with a consistency model is commonly done via knowledge distillation from a pre-trained text-to-video generation model [191, 197].

3.2.4 Decoupled Learning. Learning the text-to-video generation model inherently entails higher complexity than learning the text-to-image generation model. This is because the former needs to ensure quality output from visual content and temporal consistency aspects. To circumvent any possible trade-off between these two aspects that may occur with a single training, some models propose to decouple these aspects into two learning pathways. Follow Your Pose [134] model is among the first that implement this decoupling strategy. Motivated by the scarcity of the variety of poses depicted in the video data, the model embeds a certain pose into the output video via separate learning. This approach is also implemented by other text-to-video generation models [5, 52, 92, 122, 234]. Besides training separate pathways in one stage, other models devise different approaches, such as training the appearance module and the motion module in separate learning stages [191, 237] and learning appearance aspect through first frame synthesization and the motion aspect through subsequent frame prediction [56, 208, 228, 232].

3.2.5 Modular Generation. The burden of optimizing superior visual content and consistent temporal progression, coupled with the requirement to generate realistic high-resolution video output, has motivated many text-to-video generation models to leverage modular generation strategy. The gist of this method is to let the base model generate sparse and low-resolution outputs to alleviate the intensive computational load. Pioneered by PYoCo [55], these outputs are later refined to a higher resolution with a train of refinement models such as temporal interpolator, spatial super-resolution generator, and temporal super-resolution generator. Subsequent models [5, 12, 15, 68, 87, 113, 194, 200, 229, 231] after PYoCo follow this approach to generate realistic video output with minimum computational effort. Besides explicitly incorporating additional modules after the base generation one, another method refines the sparse frame generation to high frame rate videos through a multi-generation process [25, 85, 219].

4 DATASETS AND METRICS

4.1 Dataset

In this section, we present an analysis of the most commonly used datasets for text-to-video generation tasks. We focus on the five most frequent datasets, examining their usage, advantages, usefulness, and weaknesses in generating video from text. By exploring the strengths and limitations of these datasets, we aim to provide insights into their roles in advancing the generation techniques and highlight areas for potential improvement in future studies. A more detailed and comprehensive list of datasets is presented in the Appendix A.1.

In general, larger and more diverse datasets improve the model performance and generalization. Figure 7 illustrates this scaling and diversity law of text-to-video datasets across various applications. Smaller datasets such as Charades and Vimeo-90K focus on specific, specialized tasks, while datasets like LAION-5B, WebVid, and VAST-27M contain many pairs and are used in various applications. This indicates the significance and comprehensive scaling of datasets used to train the text-to-video generation models. Furthermore, Figure 8 illustrates the usage statistics of important datasets in main articles curated for this survey on text-to-video generation. The WebVid-2M and WebVid-10M datasets

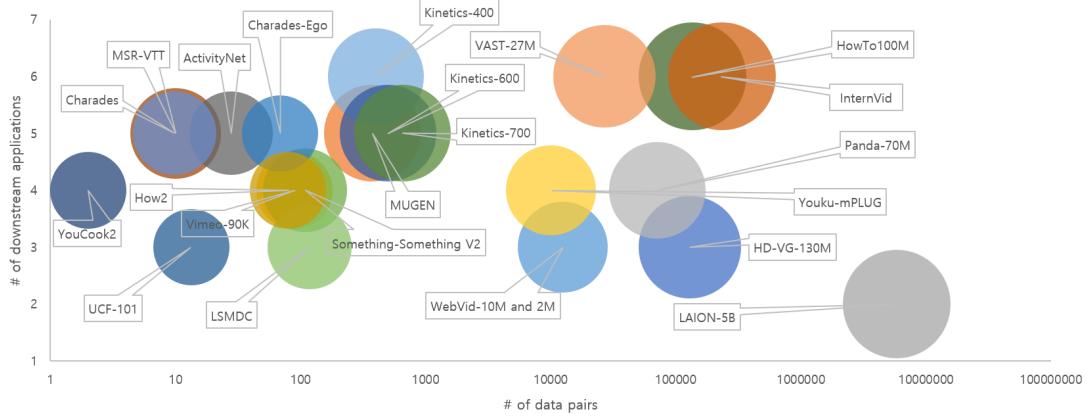


Fig. 7. Scaling and diversity of text-to-video datasets across various applications. The circle diameter indicates the dataset's magnitude.

are the most commonly used, appearing in 34 papers, followed by MSR-VTT and UCF-101, which are used in 11 and 10 papers, respectively. The chart highlights the prominence and significance of these datasets in the field, showcasing their critical role in advancing text-to-video generation techniques.

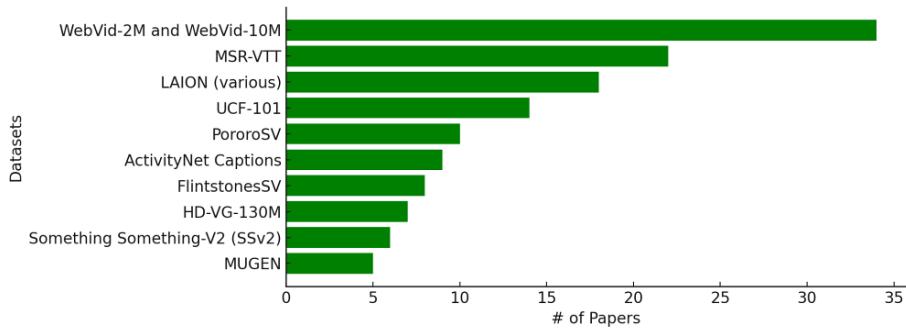


Fig. 8. Usage statistics of significant datasets in text-to-video generation research based on main papers we curated.

- **WebVid-10M and WebVid-2M [9].** The WebVid dataset, particularly WebVid-10M, is frequently used in various research projects and experiments. This large-scale dataset consists of approximately 10 million short video clips, each averaging 18 to 30 seconds in duration, and is known for its rich and diverse content spanning multiple categories such as sports, cooking, and travel. Despite its widespread use, WebVid-10M is often criticized for its low picture quality [26], with most videos having a resolution of 336×596 pixels and the presence of watermarks [199]. These characteristics present important training issues for high-quality video generation models, as they limit the models' capacity to produce high-resolution and visually pleasing results. To solve these challenges some researchers supplement WebVid-10M with additional high-quality, watermark-free video data [65]. Despite these challenges, WebVid-10M remains an important dataset for developing and benchmarking text-to-video generation models due to its extensive and well-segmented video-text pairs.

- **MSR-VTT [213].** The MSR-VTT dataset is commonly used for evaluating text-to-video generation models because it offers a large collection of 10,000 web video clips, each accompanied by around 20 natural language descriptions. It is widely used across various studies, including VideoDrafter [127], FusionFrames [5], and HiGen [155], for assessing single-scene prompt-based video generation. The dataset’s comprehensive test set includes 2,990 videos with around 20 captions each, enabling robust testing environments as demonstrated in studies like POS [132] and VideoDirectorGPT [119]. Additionally, ModelScopeT2V [192] and other models leverage MSR-VTT for zero-shot evaluation, validating performance with metrics such as FID-vid, FVD, and CLIPSIM. This dataset’s rich annotations and open-domain video captioning capabilities make it indispensable for examining the alignment and visual quality of generated videos.
- **LAION-5b [167].** The LAION dataset has been a game-changer in text-to-video generation. With its massive collection of around 5 billion text-image pairs, LAION-5B has been used to train several advanced models. For instance, VideoDrafter [127] and VideoLCM [197] leverage subsets like LAION-2B to ensure high-quality matches between visuals and text. To tackle issues like data duplication and mismatched descriptions, the deduplicated LAION-2B [103] has been particularly useful. Despite some concerns about biases and inappropriate content in LAION-400M [168], it has still contributed to improvements in models like Phenaki [188]. LAION has proven to be a common yet valuable resource in generating video from text by providing a large and diversified dataset.
- **UCF-101 [176].** The dataset serves as a standard benchmark in the field of text-to-video generation due to its diversified collection of 13,320 video clips categorized into 101 human actions. Several studies leverage UCF-101 for evaluating the video generation models [5, 68, 132], addressing its challenge of lacking native captions by incorporating text prompts from external sources like PYoCo [55]. For instance, FusionFrames [5] is evaluated on UCF-101, demonstrating its adaptability. Similarly, state-of-the-art performance in class-conditional generations on UCF-101 is achieved by diffusion models, showing the dataset’s importance in the text-to-video generation. POS [132] utilizes all 3,783 test videos for calculating metrics like FVD and IS, highlighting UCF-101’s comprehensive utility in performance evaluation. Furthermore, research aimed at realistic datasets, such as UCF-101, focuses on the complexity and annotation restrictions, as well as its importance in advancing the capability to generate video from text.
- **PororoSV [115].** The PororoSV dataset is derived from the original Pororo video QA dataset [115], and is an important resource in text-to-video generation, particularly for story visualization tasks. It consists of 15,336 description-story pairs, with each story represented by five consecutive frames, making it ideal for generating coherent sequences of images from multi-sentence paragraphs. The dataset includes nine main characters, though their appearance frequency varies, which can impact model learning and character representation [180]. Experiments utilizing PororoSV, such as those in the Make-A-Story [160] and PororoGAN [222] studies, demonstrate its challenges due to high semantic complexity and the necessity for maintaining global consistency across dynamic scenes and characters. Models like StoryGAN [115] have leveraged this dataset to benchmark improvements in visual quality and semantic relevance, although achieving real-world applicability remains an ongoing challenge due to residual visual artifacts and the complexity of accurately modeling story flow.

4.2 Metric and Evaluation

Text-to-video generation models employ evaluation metrics to measure their generation performance. Since such models involve dual modalities, text and vision, the evaluation metric is expected to judge both modalities equally.

In practice, this means measuring both visual quality and text-vision coherence. Moreover, since a video consists of interrelated images, a metric that can probe into the temporal dimension is also needed. Nevertheless, aside from these machine evaluation systems, it is common for users to judge the model output based on human perception. We present a comprehensive list of evaluation metrics in Appendix A.2.

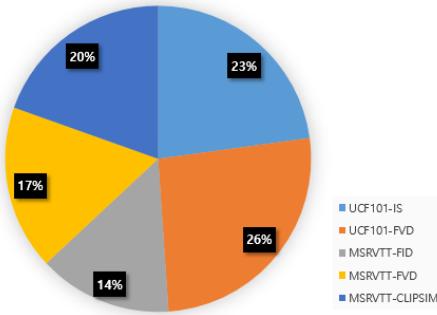


Fig. 9. Distribution of commonly used evaluation metrics (paired with the evaluation datasets) in text-to-video generation studies based on main papers we curated.

Here, we focus on the metrics most commonly used in recent literature, specifically in the main articles. Figure 9 shows how frequently different evaluation metrics are used in the studies we have conducted. The proportions highlights the prevalence of these metrics in the field. It shows five key metrics: UCF101-IS (23%), UCF101-FVD (26%), MSRVTT-FID (14%), MSRVTT-FVD (17%), and MSRVTT-CLIPSIM (20%). Each segment of the chart represents the percentage of studies that used each specific metric. This gives us a clear picture of the evaluation landscape in text-to-video generation research. Our analysis indicates that UCF101-FVD and UCF101-IS are the most commonly used metrics, underscoring their significance in assessing the performance of video generation models.

4.2.1 Visual Quality.

- **Inception Score (IS)** [164]. IS was suggested as the automatic evaluation metric to eliminate both inefficiency and inherent bias that are present in manual evaluation by humans. The evaluation is done simply by feeding all generated images into the Inception [178] network to get the conditional label distribution. IS measures the difference between the ground truth label distribution and the generated label distribution by calculating the KL divergence between the two. Although IS was originally proposed to evaluate image generation, it can also be used for video generation. IS for a video can be obtained by taking the average of ISs of all frames.
- **Fréchet Inception Distance (FID)** [80]. FID was proposed to solve the limitation of IS that prefers to use statistics of synthetic samples instead of real-world ones. The difference between ground truth samples and the generated samples distributions is measured using Fréchet distance which assumes that both distributions are Gaussian.

$$d^2((\mathbf{m}, \mathbf{C}), (\mathbf{m}_w, \mathbf{C}_w)) = \| \mathbf{m} - \mathbf{m}_w \|_2^2 + \text{Tr} \left(\mathbf{C} + \mathbf{C}_w - 2(\mathbf{C}\mathbf{C}_w)^{1/2} \right)$$

where (\mathbf{m}, \mathbf{C}) and $(\mathbf{m}_w, \mathbf{C}_w)$ are mean of Gaussian from generated and ground truth data distributions respectively. The distributions of both samples are obtained from the output of the last pooling layer of the

Inception-v3 [179] network that was pre-trained on ImageNet [40] data. Similar to IS, FID for video is obtained by taking the average of FIDs of all frames.

- **Fréchet Video Distance (FVD) [184].** FVD is the extension of FID that takes into account not only visual quality but also temporal coherence and sample diversity. Although the idea is almost similar to FID, FVD derives the feature representation of both ground truth and generated data distributions from the final layer of an inflated 3D Inception-v1 network [23]. The base architecture of the 3D Inception network was pre-trained on ImageNet, whereas the model itself was trained on the Kinetics [95] dataset. The distance between ground truth $p(X)$ and generated $q(Y)$ distributions is estimated using the maximum mean discrepancy (MMD) [64] approach to alleviate potential large errors from attempting to approximate Gaussian distributions.

$$MMD^2(q, p) = \sum_{i \neq j}^m \frac{k(x_i, x_j)}{m(m-1)} - 2 \sum_{i=1}^m \sum_{j=1}^n \frac{k(x_i, y_j)}{mn} + \sum_{i \neq j}^n \frac{k(y_i, y_j)}{n(n-1)}$$

where $x_1 \dots x_m$ and $y_1 \dots y_n$ are samples drawn from X and Y, and $k(\cdot, \cdot)$ is kernel function.

- **Generative Adversarial Metric (GAM) [89].** GAM was proposed specifically to compare the generation results of two GAN-based models. The main idea is to get two GAN models to engage in a battle that involves the exchange of generator-discriminator pairs between the two models.

4.2.2 Text-Vision Comprehension.

- **CLIP R-Precision [149].** R-precision calculates the top-R retrieval accuracy of getting the matching caption from a query image. Based on this definition, CLIP R-Precision is obtained by querying the CLIP model with the generated image and automatically checking how much the retrieved caption matches the ground truth caption. This metric is among the first attempts to measure the similarity between text and image modalities.
- **CLIP Score [79].** CLIP Score evaluates the caption-image similarity by borrowing the proficiency of the CLIP model to understand the correlation between image and text. The idea of obtaining a similarity score using this metric is rather simple. Generated image and text captions are passed through the CLIP image embedding and CLIP text embedding, respectively. The score is calculated by evaluating the cosine similarity between the two embeddings. CLIP Score for video is measured by taking the average (CLIP SIM) [205] or maximum [74] CLIP score over all frames. To minimize the influence of the CLIP model and make the evaluation metric more domain-independent, the CLIP score of the generated video can be normalized by the CLIP score of the ground truth video (CLIP RM) [205].

4.2.3 Human Perception.

- **DrawBench.** DrawBench was proposed together with Google’s Imagen [163] model that became a multidimensional text-to-image generation benchmark. The motivation behind such a benchmark is to overcome the limited visual reasoning skills and social bias of COCO [120], much like how PaintSkills [34], another evaluation benchmark, was designed. There are eleven evaluation categories in DrawBench that include color, count, spatial positioning, conflicting interaction, long description, misspelling, rare words, quoted words, and intricate prompts from DALL·E, Reddit, and DALL·E-2 preliminary evaluation [137] which are compiled in a total of 200 prompts. Nevertheless, there is also another manual human evaluation metric that is mostly utilized by models we reviewed. This metric contains components as shown in Table 2.

Table 2. Components used in human evaluation aspects.

	Visual quality	Text faithfulness	Motion realism	Temporal consistency
Make-A-Video [171]	•	•	•	
Tune-A-Video [207]		•		•
Magic Video [194]	•	•	•	•
Godiva [205]		•	•	
MMVID [74]	•		•	
CogVideo [85]	•	•	•	
StoryDALL-E [136]	•	•		•
TGANs-C [147]		•	•	•
StoryGAN [115]	•	•		•

5 APPLICATIONS AND IMPLICATIONS

5.1 Applications

5.1.1 Modeling. With the capability to understand and simulate physical worlds, it is reasonable to assume that descendant models of Sora and other advanced text-to-video generation models could perform well in 3D rendering and 3D virtual environment construction [33]. Thus, it could facilitate the development of the metaverse, offering a more dynamic, personalized, immersive user experience. The metaverse, conceived as a collective virtual shared space, merges multiple aspects of digital and augmented realities, including social networks, online gaming, augmented reality (AR), and virtual reality (VR) [106]. The metaverse thrives on the continuous creation and expansion of its virtual environments and experiences. Text-to-video generation models can potentially contribute to this aspect by enabling the rapid generation of 3D content that can populate these virtual worlds as construction of 3D objects; hence, 3D worlds are tedious and resource-intensive. This could include everything from environmental backgrounds and animated textures [7, 63, 209] to complex narrative sequences, thereby enriching the diversity and dynamism of the metaverse’s content landscape. Additionally, the potential to rapidly construct virtual 3D objects could open up new possibilities, making feasible what was once thought impossible in the future. The text-to-video generation model advancements suggest the potential for creating digital twins modeled after physical items. Other attributes of these items, like sound and tactile feedback (haptics), may be enhanced in addition to a series of images, for the sake of realistic copies of the AGI world model.

5.1.2 Spatial Computing. The main objective of video generation is to simulate spatial movement, particularly those uncommon in daily life. Metaverse simulation, robot learning, and autonomous driving are among the well-developed implementations of spatial computing that can benefit from these models.

Metaverse. The key features, as discussed previously, can also unleash the potential of the metaverse, through studying the interaction between virtual entities and human users. Constructed 3D environments may be used as testing grounds to evaluate activities that are difficult to carry out in the actual world. Some other user studies may raise ethical issues (e.g., racism or dark patterns [196]) and technological constraints (e.g., deploying a movable 100-meter building [39]). For example, user research on gathering feedback about placing enormous artifacts in a city might benefit from the text-to-video generation model-powered virtual worlds. As such, instead of spending a huge amount of time and financial burden to construct the artifacts, these models can serve as an enabling technology to assist researchers in understanding the user feedback in the mock-up or prototyping stages, with the benefits of avoiding hassle from

changing the real world configurations and thus disturbing people routine. On the other hand, currently, conducting user activities in mixed reality (MR) has technical restrictions, such as the imprecise placement of digital overlays in physical worlds. These limitations might negatively impact user experiences and skew user perception during research. Using the future generation of advanced text-to-video generation models, researchers may simulate augmented reality in virtual environments (i.e., virtual reality) to analyze user behaviors and their response to 3D user interfaces, with the premise that modern virtual reality headsets can provide high-quality video and seamless experiences.

Robotics. In robotic applications, text-to-video generation models can be an affordable platform for robots to learn manipulation actions [217]. This facilitates open-world learning by lowering the cost and effort for data collection on human demonstration which was initially performed by video recording of directed choreography with 3D motion capture [62]. Early attempts to leverage robot learning via video were made by supplying trajectory signals (e.g., flow vectors) to an action image [75]. From this, trajectory learning was improved by incorporating detailed text description [214] and robot state [206]. Such a development encourages the works in robot learning to assimilate recent generative AI models such as text-to-image [93] and text-to-video [45] generation models which inherit the power of LLM to act as policy generator or reward for reinforcement learning [46, 175]. This advancement enables the realization of scalable, unsupervised, and generalizable robot learning.

Autonomous driving. Similar to the robotic application, text-to-video generation models for autonomous driving are also mainly utilized as data generators for uncommon road scenes, such as traffic accidents [48] which are enormously costly to recreate. The availability of such data is highly beneficial for designing autonomous vehicles' safety features [38] as it enables learning the trajectory before and after the accident. Further, video generation for this field can also provide panoramic driving scene data in various road conditions that were originally limited in supply [114, 202]. Output generated by text-to-video generation models enables autonomous vehicle intelligence systems to learn directly from real-world environments instead of being confined in the game environment which was commonly used to simulate the driving scene [112, 198].

5.1.3 Media and Creative. Video content generation is one of the substantial media applications utilizing text-to-video generation models [238]. Many entrepreneurs have leveraged the text-to-video generation model to develop various media products including communication, art, and education.

Communication. Communication media is perhaps the most flourishing field of AI-generated video implementation. Currently, there are about nine text-to-video content generation tools offering professional services covering employee hiring and orientation, business presentations, news broadcasts, product commercials and marketing [153], and social media content. With these applications, users can input text prompts via various mediums including plain text, .pdf and .ppt files, and even the URL that hosts the article they want to convert into video. Figure 10 presents a few examples of AI-generated video content from DeepBrain AI¹. Other than these applications, perhaps, one of the most beneficial implementations of text-to-video generation is to generate visual alerts for disasters [204].

Creative. Moreover, text-to-video content generation is also a booming business in creative media including animation and filmmaking. Services offered by entrepreneurs in this market cover many artistic tasks such as style transfer, inpainting, color grading, motion speed editing, face blurring, scene detection, depth-of-field editing, background noise removal, subtitling and dubbing, green screening, and audio reactivity. In addition to text prompts, in this application,

¹<https://www.deepbrain.io/>



Fig. 10. Use cases of text-to-video generation applications for general communication: presidential campaign (*left*) and news broadcasting (*right*).

users can input images, paintings, or music to enhance the aesthetic of the generated video. Figure 11 presents two sample products from Runway AI² and Kaiber³ that have been used commercially.



Fig. 11. Use cases of text-to-video generation applications for creative industry: music video (*left*) and film (*right*).

Education. A further extension of text-to-video generation in the media realm is for educational purposes. This model is capable of dramatically transforming educational approaches, offering innovative strategies to enrich teaching and learning experiences. Videos have been widely discussed and applied in education [144] due to their capability to improve students' motivation [1] and self-direction [104]. Furthermore, the evolution of teaching media from text to video could provide students with deep understanding by visualizing abstract and complex concepts (e.g. science education) [187], such as the visualization of electricity flows. Therefore, the implementation of text-to-video technology in education has the potential to greatly enhance the effectiveness of instructors and their audience engagement by converting their lecture notes into video format [2]. Video-assisted education has several advantages, including the possibility for students to provide more elaborate explanations of complex ideas and attain higher levels of learning compared to traditional methods.

5.1.4 Healthcare. Despite lingering doubts about its trustworthiness, text-to-video generation holds the potential to improve the healthcare industry. Current studies primarily implement these models in healthcare education and medical imaging.

Medical training. In medical education, text-to-video generation models can help to create training videos for healthcare practitioners whose amount is considerably limited since the collection process involves real medical cases [165]. This is particularly useful in generating educational videos that involve rare cases.

²<https://runwayml.com/>

³<https://kaiber.ai/>

Medical imaging. In its function to enhance medical imaging, text-to-video generation models have been explored in various examinations such as radiology [13], CT scan [73], and endoscopy [110]. Besides, the text-to-video generation model may promote the equal distribution of health services between highly developed areas and rural areas. One interesting study by Loh and Then [126] envisions the data conversion of echocardiogram video into text to facilitate a more economical and faster information transmission between two contrastive regions. Here, the text-to-video generation model serves as a converter in the receiver end that reconstructs the echocardiogram video from the transmitted text.

5.2 Implications to World Model

5.2.1 Sora: Model That Simulates the World. With the mission of realizing AGI, OpenAI claims that its recent text-to-video generation model, Sora⁴, is a world model. This claim is rooted in Sora's ability to generate a hyper-realistic one-minute-long video which has set a seemingly rigid boundary from the existing well-known text-to-image generation models (e.g., DALL·E and Midjourney) and its peer text-to-video generation models. Upon witnessing a handful of Sora's generated video samples, the public concurs to admit its impressive generation capability in maintaining objects' 3D consistencies, temporal smoothness, and realistic physical simulation. Sora's proficiency mainly comes from two fundamental factors, model size and training data scale. Built upon diffusion transformer (DiT) [150] that unifies the expertise of transformer in handling high-dimensional data and the competence of diffusion in generating high-quality visual output, Sora can achieve the level of scalable and generalizable text-to-video generation model. Moreover, Sora collaborates with GPT-4 to enhance users' simple prompts to highly descriptive ones, potentially dictating comprehensive narratives for the video scene. Given these powerful supports, it is expected that the model that fundamentally relies on patch learning can simulate real-world movement. While being acknowledged by AI researchers, such as NVIDIA's senior scientist, Jim Fan⁵, Sora's introduction to the public raises several critics from other influential stakeholders. For instance, Yann LeCun⁶, Meta's chief AI scientist, argues that Sora is a half-cooked model. His concern comes from Sora's strong claim as a world model despite evident failures observed from its generation samples in understanding the world. He further pinpoints that Sora's primary limitation is its narrow understanding of the causal and compositional reasoning of the real world. These lukewarm receptions from leading AI scientists infer that despite text-to-video generation model's great potential in simulating the world, modeling the world requires numerous considerations beyond the model scalability. Indeed, today's world model may seem to be represented by three capabilities, visual, memory, and controller (Figure 12), which can be translated to data, architecture, and objective function, respectively [71]. Still, the generative AI that aims to simulate the world needs to implicitly learn principal components of world model [16] (i.e., theory, metaphor, analogy, policy, empirical data, stylized facts, and mathematical concepts and techniques).

5.2.2 Sora's Limitation. As discussed in the previous paragraph, Sora is possibly weakest at its understanding of real-world causal reasoning. Nevertheless, we find more shortcomings that we observe from the generated video samples. We summarise them as follows.

⁴<https://openai.com/sora>

⁵<https://x.com/DrJimFan/status/1758210245799920123>

⁶https://www.linkedin.com/posts/yann-lecun_modeling-the-world-for-action-by-generating-activity-7165252916063248384-QWwU/

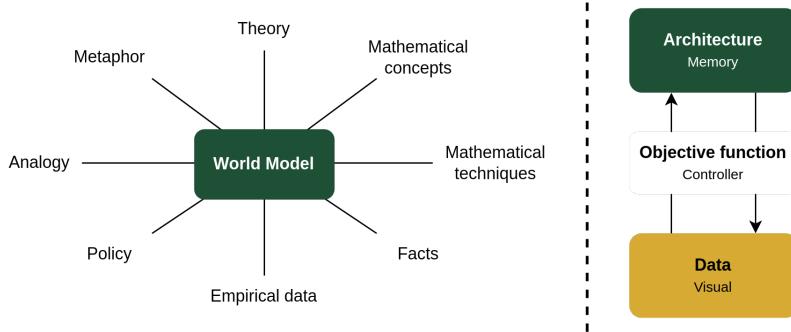


Fig. 12. Elements of world model adapted from Boumans [16] (left) and world model aspects of artificial intelligence adapted from Ha and Schmidhuber [71] (right).

Multiple Entities Handling. Generation models often perform poorly in creating a scene where several entities with similar appearances exist. Failure cases include sudden entity cloning, multiple entity dilution, and entity retraction into unrecognizable form (Figure 13).



Fig. 13. When multiple entities of similar appearance exist, video generation models often hallucinate by (a) cloning entities or (b) retracting to an unrecognizable form. Affected entities are in yellow boxes.

Causal-Effect Learning. One substantial failure of text-to-video generation models is their ineffectiveness in understanding dynamic scenes. These models have not yet been able to predict the reaction upon the occurrence of an action. Examples of cases are the inability to understand the textural relationship between interacting objects, inability to follow motion harmonization, and negligence in causal-effect ordering (Figure 14).



Fig. 14. Example of flaws in understanding causal-effect relationship; (a) liquid leaks before glass shatters and (b) candle flames do not have uniform direction and stay still despite being blown. Affected entities are in yellow boxes.

Physical Interaction. One of the reasons why video is a preferred medium to explain intricate concepts or instruction that is too tedious to be elaborated with text is because video can correctly simulate those abstract concepts in the physical world. However, synthetic video generated by text-to-video generation models is limited in simulating the proper physical interaction. This drawback includes negligence to simulate basic physics law, inability to grounding and grasp, and temporal inconsistency in displaying an object’s physical state (Figure 15).



Fig. 15. Sora’s example limitations in physical and interaction understanding include failure to understand that liquid must flow to lower ground level (*left*), the ball must not penetrate the solid ring (*middle*), and plastic chair isn’t molded from clay (*right*). Affected entities are in yellow boxes.

Scale and Proportion Understanding. Object scale and proportion are other important aspects of scene understanding. Similarly, both are also crucial factors in the video generation task. Meanwhile, proper handling of these elements is still challenging, even in a large vision model like Sora. Figure 16 illustrates a few failure cases of Sora in handling object scaling and proportions. We notice that these faults mostly happen when the scene is generated using intricate camera movement, such as rotation or changing of height from the ground level. Since video is fundamentally a stack of image frames, we conjecture that the scaling failure is caused by the transitional incoherence between frames due to the difficulty in interpreting non-linear viewpoint alteration. Nevertheless, there could exist reasons that stem from the prompting insufficiency or insufficient motion data diversity. For instance, text prompts used to generate the left and right video screen in Figure 16 are only “A beautiful homemade video showing the people of Lagos, Nigeria, in the year 2056. Shot with a mobile phone camera.” and “Beautiful, snowy Tokyo city is bustling. The camera moves through the bustling city street, following several people enjoying the beautiful snowy weather and shopping at nearby stalls. Gorgeous Sakura petals are flying through the wind along with snowflakes.”. Notice how none of the text prompts describes how camera angles should be created.

Object Hallucination. We refer to hallucination as the case where a new object suddenly appears or disappears from the generated video screen. Large text-to-vision models like Sora still suffer from this limitation. Figure 17 illustrates some failure cases caused by hallucinations. For instance, the first video scene changes from a market of the same elevation to a cityscape of a different elevation. Additionally, the second video scene illustrates how buses disappear upon occlusion by trees. From these cases, we infer that hallucination occurs when an object undergoes severe occlusion. Moreover, motion-based frame interpolation may promote the video scene to create a seemingly illogical object.

5.3 Ethical and Social Implications

The significant advancement in generative AI also entails numerous drawbacks including, but not limited to, misusage, privacy and copyright, fairness, and transparency [72].



Fig. 16. Cases of object scaling and proportion failures in Sora; dwarf-size crowds and normal-size men in one frame (*left*) and normal-size pedestrians and a giant-size couple in the same frame (*right*). Affected entities are in yellow boxes.



Fig. 17. Hallucination cases from Sora include abrupt scene change (*first row*) and emergence of new object (*second row*). Affected entities are in yellow boxes.

5.3.1 Misuse. Hyper-realistic videos generated from visual generative AI may be misused to create emotionally manipulative content that propagates misinformation during critical events such as elections [121]. The menace of disinformation through the fabricated videos depicting politicians in non-existent scenarios [53, 141], poses a significant risk of distorting public opinion. Additionally, such a model may generate fake content that threatens personal privacy and safety [142]. The most terrifying misusage of text-to-video generation models is perhaps the creation of deepfakes by the plaintiff intended for evidence in court trials which may exacerbate the pressure on the defendant side [58]. Misuse can also happen in the context of human-AI collaboration. For instance, instead of promoting the text-to-video generation model as a tool to reduce employee workload, business owners tend to consider it as an economical replacement for human labor. While this may be superficially beneficial, there are various detrimental effects, such as creativity killing and monitoring absence [145].

5.3.2 Privacy and Copyright. A common approach in massive data collection to train scalable text-to-video generation models is to scrape internet data. This data contains many personal identifications whose owner may or may not intend to spread. If the generation model suffers from a data memorization issue, someone else's face may surface in the generated video, leaking the privacy of a particular individual [131]. Further, the copyright definition in generative AI is still obscure. If someone else's work accidentally appears in the generated content, it is still uncertain whether the

user of generative AI model can be considered to infringe the original artist’s copyright. This is a dilemma because the AI user also contributes his creative thinking to designing prompts that could engineer such an artwork [131].

5.3.3 Fairness. Fairness has become a long-standing challenge in today’s generative model. Stereotyping is the most commonly found issue in any foundation models today, including vision. For instance, vision generative models like Stable Diffusion and DALL-E were found to amplify bias in gender and race [183]. The issue of fairness mainly comes from the training datasets. For the sake of simplicity, many text-to-video generation models, even generative AI models in general, were trained on data that can be mined with English descriptions. Nevertheless, this data distribution is skewed towards the Western culture which will inherently make the model generate Western-like output [54]. Although many researchers have tried to disclose this issue in some publicly available foundation models, the problem persists.

5.3.4 Transparency. Although corrective action like deepfake detection seems to be the most chosen way by policy-makers, limiting misconduct in applying generative AI models can be done preventively. Forcing the model to become more transparent can be one of the options. The transparent generative AI can be achieved, for instance, through leveraging an explainable AI system that can reveal the underlying “path” on how the user’s instruction is translated into the output video [70]. Nevertheless, the road ahead in large-scale implementation of such a measure may be full of challenges [69]. The reason is mainly for strategic purposes because disclosing the underlying mechanism of how a commercial generative product works may result in potential competition risks among business players in the same market.

6 DISCUSSION AND FUTURE DIRECTION

Despite the acclaimed success of the text-to-video generation model, the aforementioned limitations and adverse impacts are non-trivial and may trigger inhibition among the users’ community. For this reason, the research community is left with hefty homework to ensure that the generation model is indeed reliable enough to be called a world model. Here we list some suggestions inferred from our previous discussions.

6.1 Balancing Data Scaling and Class Selection

From Section 5.2.2, we can infer that simply scaling the text-to-video generation model does not guarantee that the model will be able to give near-real-world performance. Learning from an immeasurable amount of data may help the model to identify lots of real-world terms. However, that does not necessarily mean that the model also learns to perceive and fathom the knowledge beyond such data. Further, some limitations in video generation may stem from the choice of pre-training data [90]. Therefore, carefully learning the elemental distribution in the pre-training data may be one of the essential choices to scrutinize to increase the performance of such a generative model.

6.2 Automatic Evaluation for Text-Vision Alignment

Aligning text with video in a generation task is a non-trivial task. The difficulty in assessing the textual faithfulness of vision generation model output is a sign of this issue. Currently, there are only a handful of studies that implement this faithfulness evaluation system automatically. A common approach is to feed the output back to either classification models (e.g., Inception-v3 [77]) or vision-language representation models (e.g., CLIP). Another approach is to evaluate during training by the feedback verification mechanism. This process may borrow LLM (known for its reasoning capability [203]) to output a confidence score between the generated visual output and the text prompt [128]. Another approach is to leverage the video editing procedure, where the similarity score is measured between the noised output

video reconstructed using the same generation prompt and the original video output from the text-to-video generation model [221]. Nevertheless, a single text prompt can be interpreted in a hundred ways, and thus, the quality of the generated video must be evaluated from multidimensional perspectives, such as reasoning, causal effects, and spatial relationship.

6.3 Multimodal Input - Multitask Output

One of the fundamental goals of the computer vision model is to realize a general model. The general model, akin to GPT in natural language processing, is a single AI model that can process the input of various modalities and perform miscellaneous downstream tasks. The actualization of such a model in the computer vision domain, however, is still in its infancy. Only a few studies have touched upon the implementation of the general model, particularly from the perspective of text-to-video generation. In general, the model accepts text, audio, image, video, and object localization signals (e.g., bounding box, segmentation mask, depth map). Further, the model can perform diverse video-related downstream tasks, such as video generation, video editing, and video stylization. The pioneering research in this direction is CoDi [181]. Based on diffusion model architecture, CoDi efficiently handles challenges pertaining to multimodal processing (e.g., data scarcity and computational complexity) through decomposable generation. Particularly, it trains each modality-specific model individually before integrating them through latent alignment that attends to each other's modalities. The subsequent models after CoDi follow this decoupled generation concept to maintain training efficiency. For instance, VideoPoet [101] applies disintegrative input handling through modal-specific tokenizers before these tokens are handled by a decoder-only transformer that performs the generation autoregressively. Autoregressive model architecture has also been selected as the backbone of WorldDreamer [199]. WorldDreamer performs a decomposed tokenization operation similar to that of VideoPoet. The discrepancy between these two models lies in the masking strategy. While the VideoPoet only predicts the mask of the next token, WorldDreamer can perform parallel prediction thanks to its cosine scheduling dynamic masking strategy.

6.4 Human-Controlled Generation

A well-known mechanism of human-AI relationship is realized through reinforcement learning with human feedback (RLHF). Not only generating hyper-realistic output, RLHF also entails other benefits such as correcting model reasoning and safeguarding against malicious input triggers. For instance, powerful generative AI models such as GPT-4 have already been trained to reject harmful user instruction through reward learning with RLHF [20]. Unfortunately, RLHF exploration in the text-to-video generation realm is still in its infancy, with only a handful of recent works incorporating such a method. For instance, VideoDreamer [24] first lets humans choose a few of the most satisfyingly generated videos and feeds these picks back to fine-tune the generation models. Nevertheless, the research community may need to explore beyond simple RLHF as incorporating human feedback into the model may also cause unwanted consequences such as hurting the model's general capabilities [60] and triggering the model's confusion in defending its belief about the factual information [193].

6.5 Edge Generation

As video generation from text is increasingly utilized in various applications, the ability to generate with low computational resources and low latency is gradually becoming more preferred, similar to other powerful vision models [224]. For instance, generating a short video for social media content will be more convenient if performed via mobile devices [238]. Such a condition will better facilitate business owners or companies controlling customer engagement

activities. Another example of edge text-to-video generation is its implementation for MR experiences. Generating a virtual environment directly in an MR headset will enable abundant flexibility for the user, including prototyping and seamless virtual-physical interaction. Nevertheless, current text-to-video generation models still demand huge computational infrastructure to perform well. Perhaps the reason is inherent to videos, which have substantially higher-dimensional data than images and text (due to the temporal dimension).

6.6 Deepfake Control

Deepfake crafting has significantly improved with the rapid development of generative AI because such a technology promotes the democratization of content creation, lowering the barrier for novice technology adapters or low-resource users. Particularly for text-to-video generation models that generate real-world simulation, deepfake may be generated in a hyper-realistic manner. Most deepfake detection technologies rely on visual content and its impact on the public (e.g., user engagement) [17]. Although some works also attempt to integrate these elements to perform detection, such an effort may be insufficient to tackle the misuse case of the text-to-video generation model, given its realistic output. Thus, we suggest that the research community in the text-to-video generation explore other methods beyond classic approaches, such as back-tracing through the life-cycle of deepfake generation [148].

7 CONCLUSION

The arrival of Sora which can generate hyper-realistic video in the family of generative AI has surfaced the importance of a profound understanding of the underlying enabling mechanics of text-to-video generation models. Our survey pinpoints that these models are constructed upon many intricate features (e.g., temporal conditioning, efficient learning, and human feedback) that diffuse with the core building blocks, elevating their importance more than a mere expansion of text-to-image generation models. Through critical exploration centered on Sora's limitation, we also highlight that current shortcomings in text-to-video synthesis potentially arise from but not limited to the scant investigation in datasets, evaluation metrics, and human-controlled generation. These findings call for novel research directions in text-to-video generation beyond scaling up the model parameter or training data that may emerge as blue oceans for the research community. We hope that future studies that transpire from our survey can originate from various domains and solve diverse challenges in synthesizing video from text, to foster the realization of the world model.

REFERENCES

- [1] Lakmal Abeysekera and Phillip Dawson. 2015. Motivation and cognitive load in the flipped classroom: definition, rationale and a call for research. *Higher education research & development* 34, 1 (2015), 1–14.
- [2] Adebowale Jeremy Adetayo, Augustine I Enanmudu, Folashade Munirat Lawal, and Abiodun Olusegun Odunewu. 2024. From text to video with AI: the rise and potential of Sora in education and libraries. *Library Hi Tech News* (2024).
- [3] Daechul Ahn, Daneul Kim, Gwangmo Song, Seung Hwan Kim, Honglak Lee, Dongyeop Kang, and Jonghyun Choi. 2023. Story visualization by online text augmentation with context memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3125–3135.
- [4] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. 2023. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477* (2023).
- [5] Vladimir Arkhipkin, Zein Shaheen, Viacheslav Vasilev, Elizaveta Dakhova, Andrey Kuznetsov, and Denis Dimitrov. 2023. FusionFrames: Efficient Architectural Aspects for Text-to-Video Generation Pipeline. *arXiv preprint arXiv:2311.13073* (2023).
- [6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6836–6846.
- [7] Samaneh Azadi, Akbar Shah, Thomas Hayes, Devi Parikh, and Sonal Gupta. 2023. Make-an-animation: Large-scale text-conditional 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15039–15048.
- [8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

- [9] Max Bain, Arsha Nagrani, G  l Varol, and Andrew Zisserman. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 1708–1718. <https://doi.org/10.1109/ICCV48922.2021.00175>
- [10] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. 2019. Conditional GAN with Discriminative Filter Generation for Text-to-Video Synthesis.. In *IJCAI*, Vol. 1. 2.
- [11] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. 2022. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324* (2022).
- [12] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. 2024. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945* (2024).
- [13] Nathan Benaich, Elliot K Fishman, Steven P Rowe, Linda C Chu, and Elias Lugo-Fagundo. 2023. The Current State of Artificial Intelligence and Its Intersection With Radiology. *Journal of the American College of Radiology* (2023).
- [14] Andreas Blattmann, Tim Dockhorn, Sumanth Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).
- [15] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22563–22575.
- [16] Marcel Boumans. 2004. *How economists model the world into numbers*. Vol. 4. Routledge.
- [17] Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2023. Combating online misinformation videos: Characterization, detection, and future directions. In *Proceedings of the 31st ACM International Conference on Multimedia*. 8770–8780.
- [18] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the iee conference on computer vision and pattern recognition*. 961–970.
- [19] Sirley Carballo. 2024. How openai's Sora is Transforming Higher Education Video content. <https://www.enrollify.org/blog/how-openais-sora-is-transforming-higher-education-video-content>
- [20] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. 2024. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems* 36 (2024).
- [21] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A Short Note about Kinetics-600. <https://doi.org/10.48550/arXiv.1808.01340> [cs.CV]
- [22] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987* (2019).
- [23] Jo  o Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4724–4733. <https://doi.org/10.1109/CVPR.2017.502>
- [24] Hong Chen, Xin Wang, Guanning Zeng, Yipeng Zhang, Yuwei Zhou, Feilin Han, and Wenwu Zhu. 2023. Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. *arXiv preprint arXiv:2311.00990* (2023).
- [25] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. 2023. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512* (2023).
- [26] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047* (2024).
- [27] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. 2018. Neural ordinary differential equations. *Advances in neural information processing systems* 31 (2018).
- [28] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2024. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems* 36 (2024).
- [29] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. 2024. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. *arXiv preprint arXiv:2402.19479* (2024).
- [30] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. 2023. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840* (2023).
- [31] Xi Chen, Zhiheng Liu, Mengting Chen, Yutong Feng, Yu Liu, Yujun Shen, and Hengshuang Zhao. 2023. LivePhoto: Real Image Animation with Text-guided Motion Control. *arXiv preprint arXiv:2312.02928* (2023).
- [32] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2023. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*.
- [33] Zuyan Chen, Shuai Li, and Md Asraful Haque. 2024. An Overview of OpenAI's Sora and Its Potential for Physics Engine Free Games and Virtual Reality. *EAI Endorsed Transactions on AI and Robotics* 3 (2024).
- [34] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053* (2022).
- [35] Kyunghyun Cho, Bart Van Merri  nboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

- [36] François Fleuret. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1800–1807. <https://doi.org/10.1109/CVPR.2017.195>
- [37] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*. 720–736.
- [38] Minh-Son Dao, Muhamad Hilmil Muchtar Aditya Pradana, and Koji Zetsu. 2023. MM-TrafficRisk: A Video-based Fleet Management Application for Traffic Risk Prediction, Prevention, and Querying. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 1697–1706.
- [39] Jeffrey A. Delmerico, Roi Poranne, Federica Bogo, Helen Oleynikova, Eric Vollenweider, Stelian Coros, Juan Nieto, and Marc Pollefeys. 2022. Spatial Computing and Intuitive Interaction: Bringing Mixed Reality and Robotics Together. *IEEE Robotics & Automation Magazine* 29 (2022), 45–57. <https://api.semanticscholar.org/CorpusID:245982069>
- [40] Jia Deng. 2009. A large-scale hierarchical image database. *Proc. of IEEE Computer Vision and Pattern Recognition, 2009* (2009).
- [41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL* (2019).
- [42] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis, Vol. 34. 8780–8794.
- [43] Rohan Dhesikan and Vignesh Rajmohan. 2023. Sketching the future (stf): Applying conditional control techniques to text-to-video models. *arXiv preprint arXiv:2305.05845* (2023).
- [44] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. <https://doi.org/10.48550/arXiv.2010.11929> [cs.CV]
- [45] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. 2023. Video language planning. *arXiv preprint arXiv:2310.10625* (2023).
- [46] Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee, Daniar Hafner, and Pieter Abbeel. 2024. Video prediction models as rewards for reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2024).
- [47] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *CVPR*.
- [48] Jianwu Fang, Lei-Lei Li, Kuan Yang, Zhedong Zheng, Jianru Xue, and Tat-Seng Chua. 2022. Cognitive accident prediction in driving scenes: A multimodality benchmark. *arXiv preprint arXiv:2212.09381* (2022).
- [49] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. 2023. Empowering dynamics-aware text-to-video diffusion with large language models. *arXiv preprint arXiv:2308.13812* (2023).
- [50] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. 2024. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems* 36 (2024).
- [51] Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell. 2023. Tell me what happened: Unifying text-guided video completion via multimodal masked video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10681–10692.
- [52] Rinon Gal, Yael Vinker, Yuval Alaluf, Amit H Bermano, Daniel Cohen-Or, Ariel Shamir, and Gal Chechik. 2023. Breathing life into sketches using text-to-video priors. *arXiv preprint arXiv:2311.13608* (2023).
- [53] Sandro Gatra. 2024. Video kampanye “deepfake” soeharto, Pantaskah? <https://nasional.kompas.com/read/2024/01/15/10175131/video-kampanye-deepfake-soeharto-pantaskah>
- [54] Sanjana Gautam, Pranav Narayanan Venkit, and Sourojit Ghosh. 2024. From Melting Pots to Misrepresentations: Exploring Harms in Generative AI. *arXiv preprint arXiv:2403.10776* (2024).
- [55] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. 2023. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22930–22941.
- [56] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. 2023. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709* (2023).
- [57] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017. The “something something” video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*. 5842–5850.
- [58] Sam Gregory. 2023. Fortify the truth: How to defend human rights in an age of deepfakes and generative AI. , 702–714 pages.
- [59] Jiaxi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. 2023. Reuse and diffuse: Iterative denoising for text-to-video generation. *arXiv preprint arXiv:2309.03549* (2023).
- [60] Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Chen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024. Model editing can hurt general abilities of large language models. *arXiv preprint arXiv:2401.04700* (2024).
- [61] Xianfan Gu, Chuan Wen, Jiaming Song, and Yang Gao. 2023. Seer: Language instructed video prediction with latent diffusion models. *arXiv preprint arXiv:2303.14897* (2023).
- [62] Liang-Yan Gui, Kevin Zhang, Yu-Xiong Wang, Xiaodan Liang, José MF Moura, and Manuela Veloso. 2018. Teaching robots to predict human motion. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 562–567.

- [63] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5152–5161.
- [64] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. 2022. Attention mechanisms in computer vision: A survey. *Computational visual media* 8, 3 (2022), 331–368.
- [65] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Chongyang Ma, Weiming Hu, Zhengjun Zha, Haibin Huang, Pengfei Wan, et al. 2023. I2V-Adapter: A General Image-to-Video Adapter for Video Diffusion Models. *arXiv preprint arXiv:2312.16693* (2023).
- [66] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahu Lin, and Bo Dai. 2023. Sparsectr: Adding sparse controls to text-to-video diffusion models. *arXiv preprint arXiv:2311.16933* (2023).
- [67] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahu Lin, and Bo Dai. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725* (2023).
- [68] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. 2023. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662* (2023).
- [69] Mark Gurman. 2024. Apple (AAPL) shareholders reject request for AI Transparency Report. <https://www.bloomberg.com/news/articles/2024-02-28/apple-shareholders-vote-down-request-for-ai-transparency-report>
- [70] Balint Gyevnar, Nick Ferguson, and Burkhard Schafer. 2023. *Bridging the Transparency Gap: What Can Explainable AI Learn from the AI Act?* IOS Press. <https://doi.org/10.3233/fai230367>
- [71] David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122* (2018).
- [72] Thilo Hagendorff. 2024. Mapping the Ethics of Generative AI: A Comprehensive Scoping Review. *arXiv preprint arXiv:2402.08323* (2024).
- [73] Ibrahim Ethem Hamamci, Sezgin Er, Enis Simsar, Alperen Tezcan, Ayse Gulnihal Simsek, Furkan Almas, Sevval Nili Esirgun, Hadrien Reynaud, Sarthak Pati, Christian Blauthgen, et al. 2023. GenerateCT: text-guided 3D chest CT generation. *arXiv preprint arXiv:2305.16037* (2023).
- [74] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. 2022. Show me what and tell me how: Video synthesis via multimodal conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3615–3625.
- [75] Zekun Hao, Xun Huang, and Serge Belongie. 2018. Controllable video generation with sparse trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7854–7863.
- [76] Thomas Hayes, Songyang Zhang, Xi Yin, Guan Pang, Sasha Sheng, Harry Yang, Songwei Ge, Qiyuan Hu, and Devi Parikh. 2022. Mugen: A playground for video-audio-text multimodal understanding and generation. In *European Conference on Computer Vision*. Springer, 431–449.
- [77] Kaiming He, Ross Girshick, and Piotr Dollár. 2019. Rethinking imagenet pre-training. *ICCV*.
- [78] Yingqing He, Menghan Xia, Haixin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xiantao Wang, Chao Weng, Ying Shan, et al. 2023. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940* (2023).
- [79] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2021).
- [80] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).
- [81] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.
- [82] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).
- [83] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *arXiv preprint arXiv:2204.03458* (2022).
- [84] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [85] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. *arXiv preprint arXiv:2205.15868* (2022).
- [86] Yaosi Hu, Chong Luo, and Zhenzhong Chen. 2022. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18219–18228.
- [87] Hanzhou Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibeи Yang. 2024. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *Advances in Neural Information Processing Systems* 36 (2024).
- [88] Hsin-Ping Huang, Yu-Chuan Su, Deqing Sun, Lu Jiang, Xuhui Jia, Yukun Zhu, and Ming-Hsuan Yang. 2023. Fine-grained controllable video generation via object appearance and context. *arXiv preprint arXiv:2312.02919* (2023).
- [89] Daniel Jiwoong Im, Chris Dongjoo Kim, Hui Jiang, and Roland Memisevic. 2016. Generating images with recurrent adversarial networks. <https://doi.org/10.48550/arXiv.1602.05110> [cs.LG]
- [90] Saachi Jain, Hadi Salman, Alaa Khaddaj, Eric Wong, Sung Min Park, and Aleksander Mądry. 2023. A Data-Based Perspective on Transfer Learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3613–3622. <https://doi.org/10.1109/CVPR52729.2023.00352>
- [91] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. 2023. PEEKABOO: Interactive Video Generation via Masked-Diffusion. *arXiv preprint arXiv:2312.07509* (2023).
- [92] Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. 2023. Text2performer: Text-driven human video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22747–22757.

- [93] Ivan Kapelyukh, Vitalis Vosylius, and Edward Johns. 2023. Dall-e-bot: Introducing web-scale diffusion models to robotics. *IEEE Robotics and Automation Letters* (2023).
- [94] Enis Karaarslan and Ömer Aydin. 2024. Generate Impressive Videos with Text Instructions: A Review of OpenAI Sora, Stable Diffusion, Lumiere and Comparable Models. (2024).
- [95] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. <https://doi.org/10.48550/arXiv.1705.06950> [cs.CV]
- [96] Levon Khachatryan, Andranik Mousisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15954–15964.
- [97] Doyeon Kim, Donggyu Joo, and Junmo Kim. 2020. Tivgan: Text to image generation with step-by-step evolutionary generator. *IEEE Access* 8 (2020), 153113–153122.
- [98] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
- [99] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems* 28 (2015).
- [100] Ali Kóksal, Kenan E Ak, Ying Sun, Deepu Rajan, and Joo Hwee Lim. 2023. Controllable video generation with text-based instructions. *IEEE transactions on multimedia* (2023).
- [101] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. 2023. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125* (2023).
- [102] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio. 2017. Obamanet: Photo-realistic lip-sync from text. *arXiv preprint arXiv:1801.01442* (2017).
- [103] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems* 36 (2024).
- [104] Rory A Lazowski and Chris S Hulleman. 2016. Motivation interventions in education: A meta-analytic review. *Review of Educational research* 86, 2 (2016), 602–640.
- [105] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. <https://doi.org/10.1109/5.726791>
- [106] Lik-Hang Lee, Tristan Braud, Pengyuan Zhou, Lin Wang, Dianlei Xu, Zijun Lin, Abhishek Kumar, Carlos Bermejo, and Pan Hui. 2021. All One Needs to Know about Metaverse: A Complete Survey on Technological Singularity, Virtual Ecosystem, and Research Agenda. *ArXiv* abs/2110.05352 (2021). <https://api.semanticscholar.org/CorpusID:238634091>
- [107] Seungwoo Lee, Chaerin Kong, Donghyeon Jeon, and Nojun Kwak. 2023. AADiff: Audio-Aligned Video Synthesis with Text-to-Image Diffusion. *arXiv preprint arXiv:2305.04001* (2023).
- [108] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. 2020. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. *arXiv preprint arXiv:2005.05402* (2020).
- [109] Chunye Li, Liya Kong, and Zhiping Zhou. 2020. Improved-storygan for sequential images visualization. *Journal of Visual Communication and Image Representation* 73 (2020), 102956.
- [110] Chenxin Li, Hengyu Liu, Yifan Liu, Brandon Y Feng, Wuyang Li, Xinyu Liu, Zhen Chen, Jing Shao, and Yixuan Yuan. 2024. Endora: Video Generation Models as Endoscopy Simulators. *arXiv preprint arXiv:2403.11050* (2024).
- [111] Chenghao Li, Chaoning Zhang, Atish Waghwase, Lik-Hang Lee, Francois Rameau, Yang Yang, Sung-Ho Bae, and Choong Seon Hong. 2023. Generative AI meets 3D: A Survey on Text-to-3D in AIGC Era. *arXiv preprint arXiv:2305.06131* (2023).
- [112] He Li, Ruihua Han, Zirui Zhao, Wei Xu, Qi Hao, Shuai Wang, and Chengzhong Xu. 2024. Seamless Virtual Reality With Integrated Synchronizer and Synthesizer for Autonomous Driving. *IEEE Robotics and Automation Letters* (2024).
- [113] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. 2023. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398* (2023).
- [114] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. 2023. DrivingDiffusion: Layout-Guided multi-view driving scene video generation with latent diffusion model. *arXiv preprint arXiv:2310.07771* (2023).
- [115] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6329–6338.
- [116] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. 2018. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
- [117] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. 2023. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444* (2023).
- [118] Jingyun Liang, Yuchen Fan, Kai Zhang, Radu Timofte, Luc Van Gool, and Rakesh Ranjan. 2023. MoVideo: Motion-Aware Video Generation with Diffusion Models. *arXiv preprint arXiv:2311.11325* (2023).
- [119] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. 2023. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091* (2023).

- [120] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*.
- [121] Mitchell Linegar, Rafal Kocielnik, and R Michael Alvarez. 2023. Large language models and political science. *Frontiers in Political Science* 5 (2023), 1257092.
- [122] Binhui Liu, Xin Liu, Anbo Dai, Zhiyong Zeng, Zhen Cui, and Jian Yang. 2023. Dual-stream diffusion net for text-to-video generation. *arXiv preprint arXiv:2308.08316* (2023).
- [123] Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Xintao Wang, Yujiu Yang, and Ying Shan. 2023. StyleCrafter: Enhancing Stylized Text-to-Video Generation with Style Adapter. *arXiv preprint arXiv:2312.00330* (2023).
- [124] Yue Liu, Xin Wang, Yitian Yuan, and Wenwu Zhu. 2019. Cross-modal dual learning for sentence-to-video generation. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1239–1247.
- [125] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. 2024. Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. *arXiv preprint arXiv:2402.17177* (2024).
- [126] Brian CS Loh and Patrick HH Then. 2013. Cardiac echo to text conversion: Closing the urban-rural connectivity gap. In *2013 9th International Conference on Information, Communications & Signal Processing*. IEEE, 1–4.
- [127] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. 2024. VideoDrafter: Content-Consistent Multi-Scene Video Generation with LLM. *arXiv preprint arXiv:2401.01256* (2024).
- [128] Yu Lu, Linchao Zhu, Hehe Fan, and Yi Yang. 2023. FlowZero: Zero-Shot Text-to-Video Synthesis with LLM-Driven Dynamic Scene Syntax. *arXiv preprint arXiv:2311.15813* (2023).
- [129] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. 2023. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378* (2023).
- [130] Jiaxi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. 2023. GPT4Motion: Scripting Physical Motions in Text-to-Video Generation via Blender-Oriented GPT Planning. *arXiv preprint arXiv:2311.12631* (2023).
- [131] Lingjuan Lyu, C Chen, and J Fu. 2023. A pathway towards responsible AI generated content. In *Proc. 2nd tnt'l. Joint Conf. Artificial Intelligence*.
- [132] Shijie Ma, Huayi Xu, Mengjian Li, Weidong Geng, Meng Wang, and Yaxiong Wang. 2023. Optimal Noise pursuit for Augmenting Text-to-Video Generation. *arXiv preprint arXiv:2311.00949* (2023).
- [133] Wan-Duo Kurt Ma, JP Lewis, and W Bastiaan Kleijn. 2023. TrailBlazer: Trajectory Control for Diffusion-Based Video Generation. *arXiv preprint arXiv:2401.00896* (2023).
- [134] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. 2023. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186* (2023).
- [135] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2021. Improving generation and evaluation of visual stories via semantic consistency. *arXiv preprint arXiv:2105.10026* (2021).
- [136] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2022. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European Conference on Computer Vision*. Springer, 70–87.
- [137] Gary Marcus, Ernest Davis, and Scott Aaronson. 2022. A very preliminary analysis of DALL-E 2. <https://doi.org/10.48550/arXiv.2204.13807> [cs.CV]
- [138] Amir Mazahebi and Mubarak Shah. 2022. Video generation from text employing latent path construction for temporal modeling. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 5010–5016.
- [139] Rayeesa Mehmood, Runaan Bashir, and Kaiser J Giri. 2022. Text to Video GANs: TFGAN, IRC-GAN, BoGAN. In *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Vol. 1. IEEE, 1234–1239.
- [140] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. <https://doi.org/10.48550/arXiv.1906.03327> [cs.CV]
- [141] Emily Mikkelsen. 2024. North Carolina 6th district candidate Mark Walker calls video shared by Pac a “deepfake”. <https://myfox8.com/news/politics/your-local-election-hq/north-carolina-6th-district-candidate-mark-walker-calls-video-shared-by-pac-a-deepfake/>
- [142] Dan Milmo. 2024. Company worker in Hong Kong pays out £20m in Deepfake Video Call Scam. <https://www.theguardian.com/world/2024/feb/05/hong-kong-company-deepfake-video-conference-call-scam>
- [143] Gaurav Mittal, Tanya Marwah, and Vineeth N Balasubramanian. 2017. Sync-draw: Automatic video generation using deep recurrent attentive architectures. In *Proceedings of the 25th ACM international conference on Multimedia*. 1096–1104.
- [144] Michael Noetel, Shantell Griffith, Oscar Delaney, Taren Sanders, Philip Parker, Borja del Pozo Cruz, and Chris Lonsdale. 2021. Video improves learning in higher education: A systematic review. *Review of educational research* 91, 2 (2021), 204–236.
- [145] Nnamdi Chinedu Nwanyanwu and Mercy Nwanyanwu. 2021. Utilization of Artificial Intelligence in Journalism in Nigeria. *KIU Journal of Social Sciences* 7, 2 (2021), 205–212.
- [146] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj* 372 (2021).
- [147] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. 2017. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*. 1789–1798.

- [148] Yan Pang, Yang Zhang, and Tianhao Wang. 2024. VGMShield: Mitigating Misuse of Video Generative Models. *arXiv preprint arXiv:2402.13126* (2024).
- [149] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. 2021. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.
- [150] William Peebles and Saining Xie. 2023. Scalable Diffusion Models with Transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 4172–4182. <https://doi.org/10.1109/ICCV51070.2023.00387>
- [151] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
- [152] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 724–732.
- [153] AK Pradeep, Andrew Appel, and Stan Sthanunathan. 2018. *AI for marketing and product innovation: powerful new tools for predicting trends, connecting with customers, and closing sales*. John Wiley & Sons.
- [154] Bosheng Qin, Wentao Ye, Qifan Yu, Siliang Tang, and Yueteng Zhuang. 2023. Dancing avatar: Pose and text-guided human motion videos synthesis with image diffusion model. *arXiv preprint arXiv:2308.07749* (2023).
- [155] Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. 2023. Hierarchical spatio-temporal decoupling for text-to-video generation. *arXiv preprint arXiv:2312.04483* (2023).
- [156] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xiantao Wang, Ying Shan, and Ziwei Liu. 2023. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169* (2023).
- [157] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444* (2017).
- [158] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67.
- [159] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. 2022. Make-A-Story: Visual Memory Conditioned Consistent Story Generation. *arXiv preprint arXiv:2211.13319* (2022).
- [160] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. 2023. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2493–2502.
- [161] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3202–3212.
- [162] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.
- [163] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487* (2022).
- [164] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems* (2016).
- [165] Malik Sallam, Amwaj Al-Faraajat, and Jan Egger. 2024. Envisioning the Future of ChatGPT in Healthcare: Insights and Recommendations from a Systematic Identification of Influential Research and a Call for Papers. *Jordan Medical Journal* 58, 1 (2024).
- [166] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347* (2018).
- [167] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.
- [168] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021).
- [169] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. 2018. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626* (2018).
- [170] Gunnar A Sigurdsson, Gülden Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I* 14. Springer, 510–526.
- [171] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).
- [172] Aditi Singh. 2023. A Survey of AI Text-to-Image and AI Text-to-Video Generators. In *2023 4th International Conference on Artificial Intelligence, Robotics and Control (AIRC)*. IEEE, 32–36.
- [173] Xue Song, Jingjing Chen, Bin Zhu, and Yu-Gang Jiang. 2022. Text-driven Video Prediction. *arXiv preprint arXiv:2210.02872* (2022).
- [174] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency models. *arXiv preprint arXiv:2303.01469* (2023).

- [175] Sumedh Sontakke, Jesse Zhang, Séb Arnold, Karl Pertsch, Erdem Biyik, Dorsa Sadigh, Chelsea Finn, and Laurent Itti. 2024. Roboclip: One demonstration is enough to learn robot policies. *Advances in Neural Information Processing Systems* 36 (2024).
- [176] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. <https://doi.org/10.48550/arXiv.1212.0402> arXiv:1212.0402 [cs.CV]
- [177] Rui Sun, Yumin Zhang, Tejal Shah, Jiaohao Sun, Shuoying Zhang, Wenqi Li, Haoran Duan, and Bo Wei. 2024. From Sora What We Can See: A Survey of Text-to-Video Generation. (2024).
- [178] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR*.
- [179] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*.
- [180] Gábor Szűcs and Modafar Al-Shouha. 2022. Modular StoryGAN with background and theme awareness for story visualization. In *International Conference on Pattern Recognition and Artificial Intelligence*. Springer, 275–286.
- [181] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2024. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems* 36 (2024).
- [182] Andy Thomas. 2024. Sora the future of filmmaking. <https://techduffer.com/sora-the-future-of-filmmaking/>
- [183] Nitasha Tiku, Kevin Schaul, and Szu Yu Chen. 2023. Ai generated images are biased, showing the world through stereotypes ... <https://www.washingtonpost.com/technology/interactive/2023/ai-generated-images-bias-racism-sexism-stereotypes/>
- [184] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2018. *Towards Accurate Generative Models of Video: A New Metric & Challenges*. Technical Report. <https://arxiv.org/abs/1812.01717>
- [185] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *NeurIPS* (2017).
- [186] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.
- [187] Karen L Vavra, Vera Janjic-Watrich, Karen Loerke, Linda M Phillips, Stephen P Norris, and John Macnab. 2011. Visualization in science education. *Alberta Science Education Journal* 41, 1 (2011), 22–30.
- [188] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2022. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399* (2022).
- [189] Sasha Wallinger. 2024. How openai's sora impacts the future of Music Marketing. <https://www.forbes.com/sites/sashawallinger/2024/02/17/how-openais-sora-impacts-the-future-of-music-marketing/?sh=28ead1e94831>
- [190] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. 2023. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264* (2023).
- [191] Fu-Yun Wang, Zhaoyang Huang, Xiaoyu Shi, Weikang Bian, Guanglu Song, Yu Liu, and Hongsheng Li. 2024. AnimateLCM: Accelerating the Animation of Personalized Diffusion Models and Adapters with Decoupled Consistency Learning. *arXiv preprint arXiv:2402.00769* (2024).
- [192] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571* (2023).
- [193] Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2023. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? *arXiv preprint arXiv:2302.03495* (2023).
- [194] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, et al. 2024. MagicVideo-V2: Multi-Stage High-Aesthetic Video Generation. *arXiv preprint arXiv:2401.04468* (2024).
- [195] Wenjing Wang, Huan Yang, Zixi Tu, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. 2023. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874* (2023).
- [196] Xian Wang, Lik-Hang Lee, Carlos Bermejo Fernández, and Pan Hui. 2023. The Dark Side of Augmented Reality: Exploring Manipulative Designs in AR. *ArXiv abs/2303.02843* (2023). <https://api.semanticscholar.org/CorpusID:257364951>
- [197] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. 2023. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109* (2023).
- [198] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. 2023. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777* (2023).
- [199] Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. 2024. WorldDreamer: Towards General World Models for Video Generation via Predicting Masked Tokens. *arXiv preprint arXiv:2401.09985* (2024).
- [200] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. 2023. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103* (2023).
- [201] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. 2023. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942* (2023).
- [202] Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. 2023. Panacea: Panoramic and Controllable Video Generation for Autonomous Driving. *arXiv preprint arXiv:2311.16813* (2023).

- [203] Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2022. Large language models are better reasoners with self-verification. *arXiv preprint arXiv:2212.09561* (2022).
- [204] Ru-Bin Won, Ji Hoon Choi, Minji Choi, and Byungjun Bae. 2023. Segmentation-Based Masked Sampling for text-to-animated image synthesis in disaster scenarios. In *2023 14th International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 1524–1527.
- [205] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. 2021. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806* (2021).
- [206] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. 2023. Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation. *arXiv preprint arXiv:2312.13139* (2023).
- [207] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7623–7633.
- [208] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. 2023. Lamp: Learn a motion pattern for few-shot-based video generation. *arXiv preprint arXiv:2310.10769* (2023).
- [209] Zhenyu Xie, Yang Wu, Xuehao Gao, Zhongqian Sun, Wei Yang, and Xiaodan Liang. 2024. Towards detailed text-to-motion synthesis via basic-to-advanced hierarchical diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 6252–6260.
- [210] Jinbo Xing, Menghan Xia, Yong Zhang, Haixin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. 2023. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190* (2023).
- [211] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. 2023. Simda: Simple diffusion adapter for efficient video generation. *arXiv preprint arXiv:2308.09710* (2023).
- [212] Haiyang Xu, Qinghao Ye, Xuan Wu, Ming Yan, Yuan Miao, Jiabo Ye, Guohai Xu, Anwen Hu, Yaya Shi, Guangwei Xu, et al. 2023. Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks. *arXiv preprint arXiv:2306.04362* (2023).
- [213] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5288–5296.
- [214] Yuchen Xu, Li Nanbo, Arushi Goel, Zijian Guo, Zonghai Yao, Hamidreza Kasaei, Mohammadreza Kasaei, and Zhibin Li. 2023. Controllable video generation by learning the underlying dynamical system with neural ode. *arXiv preprint arXiv:2303.05323* (2023).
- [215] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. 2019. Video enhancement with task-oriented flow. *International Journal of Computer Vision* 127 (2019), 1106–1125.
- [216] Sherry Yang, Yilun Du, Bo Dai, Dale Schuurmans, Joshua B Tenenbaum, and Pieter Abbeel. 2023. Probabilistic Adaptation of Black-Box Text-to-Video Models. In *The Twelfth International Conference on Learning Representations*.
- [217] Yezhou Yang, Yi Li, Cornelia Fermuller, and Yiannis Aloimonos. 2015. Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 29.
- [218] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. 2023. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089* (2023).
- [219] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. 2023. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346* (2023).
- [220] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. 2023. Language Model Beats Diffusion–Tokenizer is Key to Visual Generation. *arXiv preprint arXiv:2310.05737* (2023).
- [221] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. 2023. InstructVideo: Instructing Video Diffusion Models with Human Feedback. *arXiv preprint arXiv:2312.12490* (2023).
- [222] Gangyan Zeng, Zhaohui Li, and Yuan Zhang. 2019. Pororogan: An improved story visualization model on pororo-sv dataset. In *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence*. 155–159.
- [223] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. 2023. Make pixels dance: High-dynamic video generation. *arXiv preprint arXiv:2311.10982* (2023).
- [224] Chaoning Zhang, Dongshen Han, Sheng Zheng, Jinwoo Choi, Tae-Ho Kim, and Choong Seon Hong. 2023. Mobilesamv2: Faster segment anything to everything. *arXiv preprint arXiv:2312.09579* (2023).
- [225] Chaoning Zhang, Chenshuang Zhang, Chenghao Li, Sheng Zheng, Yu Qiao, Sumit Kumar Dam, Mengchun Zhang, Jung Uk Kim, Seong Tae Kim, Gyeong-Moon Park, Jinwoo Choi, Sung-Ho Bae, Lik-Hang Lee, Pan Hui, In So Kweon, and Choong Seon Hong. 2023. One Small Step for Generative AI, One Giant Leap for AGI: A Complete Survey on ChatGPT in AIGC Era. *researchgate DOI:10.13140/RG.2.2.24789.70883* (2023).
- [226] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. 2023. Text-to-image Diffusion Models in Generative AI: A Survey. *arXiv preprint arXiv:2303.07909* (2023).
- [227] Chenshuang Zhang, Chaoning Zhang, Sheng Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In So Kweon. 2023. A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI. *arXiv preprint arXiv:2303.13336* (2023).
- [228] David Junhao Zhang, Dongxu Li, Hung Le, Mike Zheng Shou, Caiming Xiong, and Doyen Sahoo. 2024. Moonshot: Towards controllable video generation and editing with multimodal conditions. *arXiv preprint arXiv:2401.01827* (2024).
- [229] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. 2023. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818* (2023).

- [230] Sibo Zhang, Jiahong Yuan, Miao Liao, and Liangjun Zhang. 2022. Text2video: Text-driven talking-head video synthesis with personalized phoneme-pose dictionary. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2659–2663.
- [231] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. 2023. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077* (2023).
- [232] Yiming Zhang, Zhenling Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. 2023. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. *arXiv preprint arXiv:2312.13964* (2023).
- [233] Minglu Zhao, Wenmin Wang, Tongbao Chen, Rui Zhang, and Ruochen Li. 2024. TA2V: Text-Audio Guided Video Generation. *IEEE Transactions on Multimedia* (2024).
- [234] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. 2023. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465* (2023).
- [235] Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [236] Pengyuan Zhou, Lin Wang, Zhi Liu, Yanbin Hao, Pan Hui, Sasu Tarkoma, and Jussi Kangasharju. 2024. A Survey on Generative AI and LLM for Video Generation, Understanding, and Streaming. (2024).
- [237] Junchen Zhu, Huan Yang, Huiguo He, Wenjing Wang, Zixi Tuo, Wen-Huang Cheng, Lianli Gao, Jingkuan Song, and Jianlong Fu. 2023. Moviefactory: Automatic movie creation from text using large generative models for language and images. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9313–9319.
- [238] Junchen Zhu, Huan Yang, Wenjing Wang, Huiguo He, Zixi Tuo, Yongsheng Yu, Wen-Huang Cheng, Lianli Gao, Jingkuan Song, Jianlong Fu, et al. 2023. Mobilevidfactory: Automatic diffusion-based social media video generation for mobile devices from text. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9371–9373.

A APPENDIX

A.1 Datasets

Text-to-video generation models employ diverse datasets for training and evaluation as different models may be constructed for different specialized downstream applications. Here, we provide a comprehensive list of these datasets.

- **WebVid-10M and WebVid-2M** [9] Both datasets contain a total of 12.5 million video-text pairs gathered from the internet. WebVid-10M contains 10 million pairs, while WebVid-2M contains 2.5 million video-text pairs. The video clips had a total length of 18 seconds, providing a thorough visual background. The descriptions had an average of 12 words; the description is an Alt-text from web photos that were found by the retrieval function.
- **MUGEN** [76] is one of the first large-scale collections of video-audio-text data that was designed explicitly for multimodal understanding and generation and was created using the CoinRun platform game. The dataset consists of 375,000 video clips, each lasting 3.2 seconds, and is paired with dense annotations, include high-quality semantic maps, and both manually collected and automatically generated text descriptions. Reinforcement learning agents are used to play the CoinRun game, resulting in a wide variety of interactions in video. In addition, the dataset has synchronized audio that was specifically designed to be played in an immersive gaming environment. The major annotation in MUGEN is video-audio-text alignments and semantic mapping, which makes it one of the most valuable for work in text-to-video generation and similar tasks.
- **LAION-5B** [167] dataset consists of more than 5.85 billion CLIP-filtered image-text pairs, of which 2.32 billion are in English. Although designed for image-text tasks, the rich and diverse content of the data is still beneficial when used to pre-train text-to-video generation methods. The dataset helps the model understand and generate video content based on text, which is a valuable prerequisite for potential further fine-tuning on video data only.
- **Vimeo-90K** [215] dataset is an extensive high-quality video dataset for low-level video processing tasks, such as frame interpolation, video denoising, deblocking, and super-resolution. It contains 89,800 video clips, where all frames have a resolution of at least 720p. The video clips are spatio-temporally matched, well-focused, and have high-quality in visual features, downloaded from Vimeo. It is split into three benchmarks designed for three video processing tasks. These benchmarks enable different developing techniques in one line, promoting comprehensive technological improvements. It has significantly improved the quality and size of previous ones, making Vimeo-90K competitive.
- **HD-VG-130M** [195] is an open-source dataset. It is designed for the task of generating high-quality text-to-video and consists of 130 million text-video pairs from various open-domain domains, meeting the need for high-definition, widescreen, and watermark-free aspects. Therefore, this dataset boosts the ability to develop video generation models with the help of plentiful resources of data refined quantitatively and qualitatively to substantially alter model output. Videos are approximately 20 seconds in length, and captions are roughly 10 words in length since they are produced by sophisticated captioning methods that accurately describe the video.
- **Something-Something V2** [57] is a large-scale dataset explicitly created for visual common sense and object interaction understanding research. It contains over 108,499 videos for 174 different classes. The length of each video is between 2 and 6 seconds. This dataset is organized based on user-generated videos and every video has a label based on a template captioned under the title of “something-something.” This template caption by viewing and drawing inferences of human common sense is designed to help. This kind of structure/dataset labeling is intended to enable the development of models that understand and predict subtleties in physical and common sense interactions, which is essential for meaningful complex scene understanding. This structure

aims to develop models capable of understanding nuanced physical interactions, crucial for complex scene comprehension and activity recognition in videos.

- **UCF-101** [176] dataset contains 101 action classes over 13,320 video clips, which were produced for action recognition. The dataset covers 27 hours of video data and depicts a large number of human actions recorded in uncontrolled settings, which include varied camera motions and cluttered backgrounds. Because the acquired data are diverse and difficult to navigate, it is a good resource for developing and evaluating action recognition algorithms.
- **MSR-VTT** [213] is a large-scale video description dataset. It contains 10,000 web video clips for a total of 41.2 hours of raw image footage, paired with about 20 sentences each, representing a total of over 200,000 sentences, all produced with the useful contribution of approximately 1,300 AMT workers. The dataset is divided into 20 different classes, offering a massive variation of sentences and vocabulary. Therefore, it is an excellent resource for training and comparing video captioning systems as well as equipment learning models connected with video-to-text translation.
- **ActivityNet** [18] dataset is a large-scale video benchmark, which includes approximately 27,800 untrimmed videos from 203 diverse activity classes, averaging 137 videos per class, collectively providing around 849 hours of video content. It is unique for its depth in activity categorization, covering a wide range of complex human activities relevant to daily life. In video in this dataset is annotated with multiple activity instances, which improves its utility for training and evaluating models across various computer vision tasks, including activity detection and classification.
- **Epic-Kitchens** [37] is designed for a better understanding of human-object interaction. The recording of the dataset captures first-person routines of kitchen-related activities. It contains more than 55 hours of video captured by 32 participants of 10 nationalities, performing unprompted routines in several kitchen environments. There are 454,300 object bounding boxes and 39,600 action tracks in it. The dataset is useful for object interaction and action prediction
- **YouCook2** [235] dataset consists of 2,000 YouTube videos related to cooking. spanning 176 hours nearly equally distributed over 89 recipes across four major cuisines. The dataset is created to help in understanding and segmenting complex cooking activities, each video is paired with detailed English sentences describing cooking steps. The annotations shows the start and end times of each procedural segment, which make the dataset important resource for developing and benchmarking video understanding models, particularly for instructional content in the culinary domain.
- **HowTo100M** [140] dataset is gathered from 1.22 million narrated instructional videos, including 136 million video clips and more than 15 years of video content, contains around 23,000 unique visual tasks such as cooking, crafts, and repairs. This dataset enables the development of a text-to-video retrieval system, needed to improve action localization algorithms through advanced video-language model training.
- **Kinetics** [95] dataset is a collection of 400 human action classes with at least 400 video clips ranging from about 10 seconds, recorded from distinct YouTube videos. The dataset is used to advance video understanding and action recognition, which includes various human action categories and subcategories, such as human-object interactions and human-human interactions, like playing musical instruments or shaking hands.
- **VAST-27M** [28] is a large-scale omni-modality video dataset that contains 27 million video clips, each including 11 captions: 5 vision captions, 5 audio captions, and 1 vision-audio-subtitle integrated captions. The captions

are produced in various types, enabling models for more intricate multi-modal tasks like video-text retrieval, captioning, and question-answering.

- **Panda-70M [29]** is a large-scale video dataset that consists of 70.8 million video clips. Each video is paired with a caption, averaging 13.2 words in narration, and comes from high-resolution, long videos to guarantee the richness and semantic coherence of the clip without any watermark. With its automatic annotation method based on multimodal data inputs, Panda-70M has many uses in video understanding, including text-driven video synthesis, video-text retrieval, and video captioning. It offers useful tools for advancing the multimodality and data efficiency of machine learning models that use visual and language data.
- **Youku-mPLUG [212]** is a large-scale Chinese video-language dataset that contains 10 million video-text pairs. Each pair averages 54.2 seconds. The dataset is sourced from 400 million raw videos from Youku, a known Chinese video-sharing website. This dataset supports many tasks, such as cross-modal retrieval, descriptive subtitle, video captioning, and video category classification. Through this dataset, the gap in Chinese video-language pre-training is reduced to support more deep-learning studies in multi-modality.
- **Charades [170]** is a data collection of 9,848 videos. Each video is a record of every 30.1 seconds of action, demonstrating 267 people from three continents. It gathered 27,847 video descriptions, 66,500 temporal-bound intervals among 157 action classes, and 41,104 labels among 46 object classes. This dataset is unique and highly efficient for the task of object detection, human-type detection, and action recognition because it covers the daily activity around the house. The Charades dataset is collected based on various considerate formations known as Hollywood in Homes. It is a novel approach where the video is recorded through lifestyle and crowdsourcing while regular people in their houses are asked to act out the prewritten scripts.
- **LSMDC [161]** is a well-comprehensible video dataset that combines Descriptive Video Service (DVS) and movie scripts, which are composed and aligned with full-length HD movies. The dataset consists of around 68k video clips and sentences sourced from 94 movies. The recorded clip and its paired caption consist of an average of 7.0 words and 4.8 seconds. The dataset can easily be understood by the models due to its many captions for a single video. This naturally helps the model to learn the plots, human interactions, and the semantics of the captions and the video.
- **Charades-Ego [169]** In this dataset, 68,536 activity instances recorded during 68.8 hours of egocentric video. There are also 66,500 activity instances of third-person video in 82.3 hours. In total, 8,000 videos revealing paired first- and third-person perspectives, provide scope for more advanced work in egocentric video classification. Charades-Ego is a database of more than 364 pairs, encompassing 31.2-second-long videos.
- **InternVid [201]** is a dataset with 234 clips that are randomly chosen from over 7 million videos equaling 760,000 hours of video in total. All InternVid video clips are described as an average of 17.6 words, and each clip is accompanied by a specific caption that explains 16 diverse settings and about 6,000 various gestures.
- **DAVIS [152]** dataset contains 50 top-quality Full HD video sequences, and 3455 annotated frames, with each video featuring pixel-accurate, per-frame ground truth segmentation. These videos are specifically recorded to cover a variety of classic video object segmentation problems, including fast motion, occlusions, and appearance changes. Each clip duration in the collection is between two and four seconds overall.
- **How2 [166]** collection comprises roughly 79,114 videos covering a wide range of instructional topics. There are almost 2,000 hours of total video content, with an average clip length of 90 seconds. The Portuguese translations of the English subtitles are included as an additional feature. The dataset was specially made to support research on multimodal language interpretation.

- **Kinetics-600 [21]** dataset is composed of around 500K action-class-aligned video clips drawn from a diverse range of activities. With at least 600 video segments in each class, the total amount of video content is enormous. Each clip, which has an average length of 10 seconds and is compiled from 600 distinct action classes on YouTube, offers an extensive amount of data for training algorithms to identify human actions.
- **Kinetics-700 [22]** dataset consists of approximately 650,000 video clips classified into 700 action classes. Each class contains no fewer than 600 videos. This version of the Kinetics dataset is an extension of Kinetics-600, adding 100 new classes while retaining almost all of the original ones from existing videos.

A.2 Metrics

As text-to-video generation models are developed upon diverse combinations of enabling technologies, they are evaluated with various metrics. Here, we provide a comprehensive list of these metrics.

A.2.1 Generative Model Evaluation Metrics.

- **Generative Adversarial Metric (GAM):** This metric assesses the discriminator's ability in a generative adversarial network (GAN) to distinguish between real and generated videos. The evaluation can be quantified by the discriminator's classification accuracy, defined mathematically as:

$$GAM = \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))]$$

where $D(x)$ is the discriminator's estimate of the probability that real data instance x is real, $G(z)$ is the generator's output when given noise z , and p_{data} and p_z are the data and noise distributions, respectively.

- **Negative Log Likelihood (NLL):** Used to measure how well a generative model predicts a sample from the data distribution, reflecting the model's accuracy:

$$NLL = -\log p_{\text{model}}(x)$$

where $p_{\text{model}}(x)$ is the probability assigned by the model to the true data point x .

A.2.2 Accuracy Metrics.

- **Classifier Accuracy:** Represents the percentage of correct predictions made by the model over a test dataset, often used to evaluate discriminative models:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

- **Classification Confusion Matrix:** Provides a visualization of the performance of an algorithm. Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class. The matrix C where element $C_{i,j}$ is the number of observations known to be in group i but predicted to be in group j .

A.2.3 Contextual and Human-Centric Metrics.

- **Contextual Consistency Metrics:** Evaluate the consistency of generated content with contextual information, commonly using metrics such as the Contextual FID:

$$CFID = \frac{1}{N} \sum_{i=1}^N \left(\mu_{\text{generated},i} - \mu_{\text{context},i} \right)^2$$

where $\mu_{\text{generated},i}$ and $\mu_{\text{context},i}$ are feature vectors of the generated image and the context image respectively.

- **Human Evaluation:** Direct assessment from human observers, often quantified through scales like MOS (Mean Opinion Score):

$$MOS = \frac{1}{N} \sum_{i=1}^N s_i$$

where s_i represents the score given by the i -th evaluator.

- **Attribute Classification Accuracy:** Measures the accuracy of attributes detected in generated videos compared to ground truth, calculated as:

$$ACA = \frac{\text{Number of correctly classified attributes}}{\text{Total attributes}}$$

A.2.4 Frame and Video Level Metrics.

- **Frame-level FID (Fréchet Inception Distance):** Measures the distance between feature vectors of real and generated frames:

$$FID_{\text{frame}} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})$$

where μ_r, μ_g are the feature means of real and generated frames, and Σ_r, Σ_g are their covariances.

- **Video-level FID:** Aggregates frame-level FIDs across a video to measure overall video quality:

$$FID_{\text{video}} = \frac{1}{T} \sum_{t=1}^T FID_{\text{frame},t}$$

- **Frame Inception Score (Frame-IS) and Video Inception Score (Video-IS):** Evaluate the clarity and diversity of generated frames and videos:

$$IS_{\text{frame}} = \exp(E[\text{KL}(p(y|x)||p(y))])$$

$$IS_{\text{video}} = \exp\left(\frac{1}{T} \sum_{t=1}^T E[\text{KL}(p(y_t|x_t)||p(y_t))]\right)$$

- **SSIM (Structural Similarity Index Measure):** Compares the similarity between two images or frames:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

- **CLIP Similarity (SIM):** Measures the semantic similarity between text and video content using CLIP embeddings:

$$SIM = \frac{v_{\text{text}} \cdot v_{\text{video}}}{\|v_{\text{text}}\| \cdot \|v_{\text{video}}\|}$$

- **Relative Matching (RM) Metric:** Evaluates the relevance of video segments to the corresponding text description:

$$RM = \frac{\sum_{i=1}^N \mathbf{1}_{\text{match}_i}}{N}$$

where $\mathbf{1}_{\text{match}_i}$ is an indicator function returning 1 if the i -th segment matches the description, 0 otherwise.

A.2.5 Human-Centric and Semantic Metrics.

- **Visual Realisticity (VR):** Assesses the photorealistic quality of generated videos, quantifying how indistinguishable they are from real-world videos. This can be measured using a perceptual realism score:

$$VR = \frac{1}{N} \sum_{i=1}^N \text{human_score}(x_i)$$

where x_i are the generated videos and human_score represents scores from human evaluators.

- **Semantic Consistency (SC):** Measures the semantic alignment between the generated video and the input text, often quantified using natural language processing tools to compare descriptions:

$$SC = \frac{1}{N} \sum_{i=1}^N \text{semantic_similarity}(x_i, t_i)$$

where t_i is the text description corresponding to the video x_i .

- **Video Captioning Accuracy:** The accuracy of captions generated automatically for videos, reflecting the relevance and correctness of the content described:

$$VCA = \frac{\text{Number of correct captions}}{\text{Total number of captions generated}}$$

- **Discriminative Evaluation:** Uses discriminative models to classify or differentiate between generated and real videos:

$$DE = \frac{\sum_{i=1}^N \mathbf{1}(\text{pred}_i == \text{real}_i)}{N}$$

where $\mathbf{1}$ is an indicator function, and pred_i is the prediction for the i -th sample.

- **R-Precision:** Measures the relevance of retrieved videos to a query in a retrieval task:

$$\text{R-Precision} = \frac{\text{Number of relevant videos retrieved}}{\text{Total number of relevant videos}}$$

A.2.6 Quantitative Performance Metrics.

- **Mean-Squared Error (MSE):** Quantifies the average of the squares of the errors between predicted and true values, important for regression tasks:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

where Y_i are actual values and \hat{Y}_i are predicted values.

- **Peak Signal to Noise Ratio (PSNR):** Measures the ratio between the maximum possible power of a signal and the power of corrupting noise:

$$PSNR = 20 \cdot \log_{10} \left(\frac{\text{MAX}_I}{\sqrt{MSE}} \right)$$

- **Learned Perceptual Image Patch Similarity (LPIPS):** Evaluates the perceptual difference between two images or videos using deep learning features:

$$LPIPS = \sum_l \frac{1}{H_l W_l} \sum_{h,w} 1 - \cos(\phi_l(x)_{hw}, \phi_l(y)_{hw})$$

where ϕ_l denotes features extracted from layer l , and H_l, W_l are dimensions at that layer.

- **Precision-Recall Distribution (PRD):** Compares the precision and recall rates of different models:

$$PRD = (\text{precision}(\theta), \text{recall}(\theta)) \text{ for thresholds } \theta$$

- **Character Classification F1 Score:** Harmonic mean of precision and recall for character recognition tasks in videos:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

A.3 Models

The following table presents a comprehensive list of 97 text-to-video generation models curated with the PRISMA framework.

Table A1. List of text-to-video generation models reviewed in our survey (in the order of publication time).

Model	Vision Processor	Language Interpreter	Temporal Handler
TGANs-C [147]	GAN	RNN	RNN
Sync-DRAW [143]	VAE	RNN	RNN
ObamaNet [102]	GAN	RNN	RNN
T2V [116]	GAN	RNN	RNN
TFGAN [10]	GAN	RNN	RNN
CMDL [124]	GAN	RNN	RNN
StoryGAN [115]	GAN	RNN	RNN
TivGan [97]	GAN	RNN	RNN
Improved-StoryGAN [109]	GAN	RNN	RNN
Godiva [205]	VQ-VAE	Transformer	Temporal attention
DuCo-StoryGAN [135]	GAN	RNN	RNN
MMVID [74]	VQ-GAN	Transformer	Temporal attention
CoGVideo [85]	VQ-VAE	Transformer	Temporal attention
Modular StoryGAN [180]	GAN	RNN	RNN
Make-a-video [171]	Diffusion	Contrastive	Pseudo-3D convolution
StoryDALL-E [136]	VQ-VAE	TF, contrastive	Temporal attention
Phenaki [188]	Autoregressive TF	Transformer	Temporal attention
TVP [173]	GAN	Transformer	RNN
[138]	GAN	Transformer	RNN
Text2Video [230]	GAN	Transformer	RNN
MAGE [86]	VQ-VAE	Transformer	Temporal attention
Follow your pose [134]	Diffusion	Contrastive	Pseudo-3D convolution
GPT4Motion [130]	Diffusion	Contrastive	LLM
CVGI [100]	GAN	CNN	RNN
Dysen-VDM [49]	Diffusion	Contrastive	Pseudo-3D convolution
Nuwa-XL [219]	Diffusion	Contrastive	Pseudo-3D convolution
Seer [61]	Diffusion	Contrastive	Pseudo-3D convolution

Continuation of Table A1

Model	Vision Processor	Language Interpreter	Temporal Handler
Text2Video-Zero [96]	Diffusion	Contrastive	Temporal attention
Tune-a-video [207]	Diffusion	Contrastive	Pseudo-3D convolution
TiV-ODE [214]	VQ-VAE	Transformer	Temporal attention
Latent-Shift [4]	Diffusion	Transformer	Temporal attention
Text2Performer [92]	VQ-VAE	Transformer	Temporal attention
AADiff [107]	Diffusion	Contrastive	Temporal attention
ControlVideo [231]	Diffusion	Contrastive	Pseudo-3D convolution
Gen-L-Video [190]	Diffusion	Contrastive	Pseudo-3D convolution
PYoCo [55]	Diffusion	TF, contrastive	Pseudo-3D convolution
STF [43]	Diffusion	Contrastive	Temporal attention
VideoFactory [195]	Diffusion	Contrastive	Temporal attention
MovieFactory [237]	Diffusion	Contrastive	Pseudo-3D convolution
Video Adapter [216]	Diffusion	Transformer	Pseudo-3D convolution
MMVG [51]	VQ-GAN	Contrastive	RNN
Animate-a-Story [78]	Diffusion	Contrastive	Pseudo-3D convolution
AnimateDiff [67]	Diffusion	Contrastive	Temporal attention
Dancing Avatar [154]	Diffusion	Transformer	LLM
DragNUWA [218]	Diffusion	Contrastive	Temporal attention
ModelScopeT2V [192]	Diffusion	Contrastive	Pseudo-3D convolution
SimDA [211]	Diffusion	Contrastive	Pseudo-3D convolution
CMOTA [3]	VQ-VAE	Transformer	RNN
LaVie [200]	Diffusion	Contrastive	Pseudo-3D convolution
LVD [117]	Diffusion	TF, contrastive	Pseudo-3D convolution
VidRD [59]	Diffusion	Contrastive	Pseudo-3D convolution
VideoDirectorGPT [119]	Diffusion	TF, contrastive	Pseudo-3D convolution
VideoGen [113]	Diffusion	Contrastive	Pseudo-3D convolution
DynamiCrafter [210]	Diffusion	Contrastive	Pseudo-3D convolution
FreeNoise [156]	Diffusion	Contrastive	Pseudo-3D convolution
LAMP [208]	Diffusion	Contrastive	Pseudo-3D convolution
MotionDirector [234]	Diffusion	Contrastive	Pseudo-3D convolution
SEINE [32]	Diffusion	Contrastive	Pseudo-3D convolution
Show-1 [229]	Diffusion	TF, contrastive	Pseudo-3D convolution
VideoCrafter1 [25]	Diffusion	Contrastive	Pseudo-3D convolution
LiveSketch [52]	Diffusion	Contrastive	Temporal attention
Emu Video [56]	Diffusion	TF, contrastive	Pseudo-3D convolution
FlowZero [128]	Diffusion	TF, contrastive	LLM
PixelDance [223]	Diffusion	Contrastive	Pseudo-3D convolution
Make-a-story [159]	Diffusion	TF, contrastive	Temporal attention

Continuation of Table A1

Model	Vision Processor	Language Interpreter	Temporal Handler
MoVideo [118]	Diffusion	Contrastive	Pseudo-3D convolution
POS [132]	Diffusion	TF, contrastive	Pseudo-3D convolution
SparseCtrl [66]	Diffusion	Contrastive	Pseudo-3D convolution
Stable Video Diffusion [14]	Diffusion	Contrastive	Pseudo-3D convolution
VideoDreamer [24]	Diffusion	Contrastive	Temporal attention
Video LDM [15]	Diffusion	Contrastive	Pseudo-3D convolution
Control-a-Video [30]	Diffusion	Contrastive	Pseudo-3D convolution
DSDN [122]	Diffusion	Contrastive	Pseudo-3D convolution
FACTOR [88]	Autoregressive TF	TF, contrastive	Temporal attention
FusionFrames [5]	Diffusion	Contrastive	Pseudo-3D convolution
HiGen [155]	Diffusion	Contrastive	Pseudo-3D convolution
I2V-Adapter [65]	Diffusion	Contrastive	Pseudo-3D convolution
InstructVideo [221]	Diffusion	Contrastive	Pseudo-3D convolution
LivePhoto [31]	Diffusion	Contrastive	Pseudo-3D convolution
Peekaboo [91]	Diffusion	TF, contrastive	Pseudo-3D convolution
W.A.L.T. [68]	Autoregressive TF	Transformer	Temporal attention
PIA [232]	Diffusion	Contrastive	Pseudo-3D convolution
StyleCrafter [123]	Diffusion	Contrastive	Pseudo-3D convolution
TrailBlazer [133]	Diffusion	Contrastive	Pseudo-3D convolution
VideoLCM [197]	Diffusion	Contrastive	Pseudo-3D convolution
VideoPoet [101]	Autoregressive TF	Transformer	Temporal attention
MagicVideo-V2 [194]	Diffusion	Contrastive	Pseudo-3D convolution
Moonshot [228]	Diffusion	Contrastive	Temporal attention
VideoCrafter2 [26]	Diffusion	Contrastive	Pseudo-3D convolution
VideoDrafter [127]	Diffusion	TF, contrastive	Pseudo-3D convolution
AnimateLCM [191]	Diffusion	Contrastive	Pseudo-3D convolution
Lumiere [12]	Diffusion	Contrastive	Pseudo-3D convolution
SceneScape [50]	Diffusion	Contrastive	Temporal attention
Free-bloom [87]	Diffusion	TF, contrastive	LLM
CoDi [181]	Diffusion	Contrastive	Pseudo-3D convolution
TA2V [233]	VQ-GAN	Transformer	Pseudo-3D convolution
WorldDreamer [199]	VQ-GAN	Transformer	Pseudo-3D convolution

End of Table