Chinese Journal of Electronics
vol.04, no.04, pp.1– 22, 2022
https://doi.org/123.123.123.123

Review

A Comprehensive Survey on Text-to-Video Generation

Firstname1 Middlename1 Lastname1[1,2,3], Firstname2 Middlename2 Lastname2[2], Firstname3 Middlename3 Lastname3[4], Firstname4 Middlename4 Lastname4[1], Firstname5 Middlename5 Lastname5[2,4], Firstname6 Middlename6 Lastname6[1], and Firstname7 Middlename7 Lastname7[4]

1. *Institution Name, City name and postal code, Country*
2. *Institution Name, City name and postal code, Country*
3. *Institution Name, City name and postal code, Country*
4. *Institution Name, City name and postal code, Country*

Corresponding author: Firstname2 Middlename2 Lastname2; Email: xxxxxxxxxx@123.com.
Received March 22, 2022; Accepted March 22, 2022; Published March 22, 2022.

Abstract — Since the release of Sora, the Text-to-Video (T2V) generation has brought profound changes to AI-generated content. T2V generation aims to generate high-quality videos based on a given text description, which is challenging due to the lack of large-scale, high-quality text-video pairs for training and the complexity of modeling high-dimensional video data. Although there have been some valuable and impressive surveys on T2V generation, these surveys introduce approaches in a relatively isolated way, lack the development of evaluation metrics, and lack the latest advances in T2V generation since 2023. Due to the rapid expansion of the field of T2V generation, a comprehensive review of the relevant studies is both necessary and challenging. This survey attempts to connect and systematize existing research in a comprehensive way. Unlike previous surveys, this survey paper reviews nearly ninety representative T2V generation approaches and includes the latest method published on March 2024 from the perspectives of model, data, evaluation metrics, and available open-source. It may help readers better understand the current research status and ideas and have a quick start with accessible open-source models. Finally, the future challenges and method trends of T2V generation are thoroughly discussed.
Keywords — Survey, Text-to-Video Generation, Generative AI, Sora Model, AIGC.

I. Introduction

Artificial Intelligence Generated Content (AIGC) is developing rapidly and has become one of the most popular topics in AI. The generative modalities of AIGC include image[1–3], video[4–6], audio[7–9], and more. We counted the number of papers published on different generated modalities in the past five years (2019 to 2023) in Figure 1. As illustrated in Figure 1(a), text-to-image (T2I) generation research has dominated the AIGC field for many years. Nevertheless, we can also see from Figure 1(b) that the development of text-to-video (T2V) generation has exploded in recent years, which may fundamentally shift the research emphasis in the future. We can see that T2I generation started early and is the focus of research. Although T2V generation started relatively late, the right graph in Figure 1 shows that it has grown rapidly in recent years.

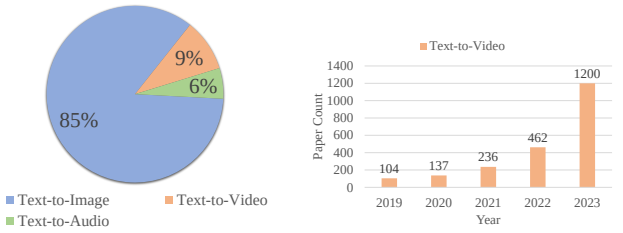


Figure 1 AIGC developments in the last five years, including Text-to-Image, Text-to-Video and Text-to-Audio.

T2V generation aims to generate high-quality videos based on a given text description, and the videos generally contain 16 frames with a duration of two seconds. It is challenging for two reasons: First, there is a lack of large-scale, high-quality text-video pairs for training; for example, tens of millions of paired data are usually required. Second, the complexity of modeling high-dimensional video data is high because 1) The semantic space for the text is much smaller than the generation space for the video frame. 2) Correct retention

of semantics and continuity between frames are required. 3) The computation power is demanding, training a T2V model like InternVid[10] typically requires 64 NVIDIA GPUs for three days.

The release of Sora[11] this year has profoundly pushed the frontier of the T2V generation. Prior to that, both academia and industry put a great effort into improving T2V generation models due to the wide application prospects. At this point, a comprehensive review of the relevant studies is both necessary and challenging. Although there have been some valuable and impressive surveys on T2V generation, these surveys introduce approaches in a relatively isolated way, lack the development of evaluation metrics, and lack the latest advances in T2V generation since 2023. Unlike previous surveys, this survey paper reviews nearly ninety representative T2V generation approaches and includes the latest method published on March 2024 from the perspectives of model, data, evaluation metrics, and available open-source.

Our survey is illustrated in Figure 2, and the organization is as follows: Section 2 clarifies the differences between this survey and others. Section 3 explores existing methods and reviews their strengths and weaknesses. Section 4 introduces current T2V datasets, while Section 5 reviews the development of metrics for evaluating T2V generation. Section 6 provides the results of the experiment on representative methods. Section 7 discusses challenges and future trends, and the last section concludes this review.

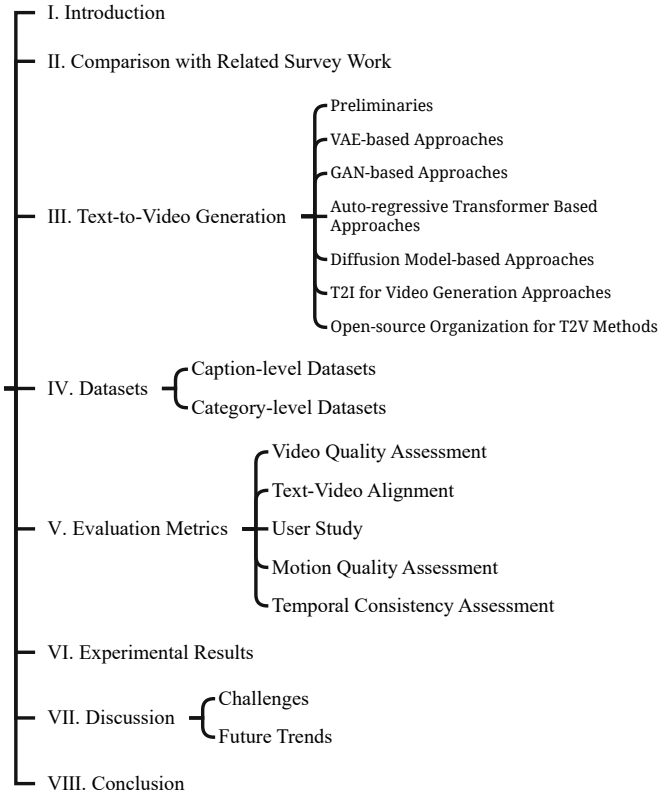


Figure 2 The organization of our survey.

II. Comparison with related survey work

Table 1 Compare our survey with existing surveys.

Survey	#Methods	Latest Pub. Year	#T2V Datasets	#Metrics
[12]	6	Oct.2022	NA	NA
[13]	16	Dec.2022	15	5
[14]	28	Oct.2023	31	4
Ours	88	Mar.2024	40	20

Table 1 presents the differences between this survey and the existing surveys. Unlike previous surveys, this survey paper reviews nearly ninety representative T2V generation approaches and includes the latest method published in March 2024. Also, more T2V datasets and metrics are comprehensively reviewed.

Singh[12] presents and compares popular T2I and T2V generation methods, discussing their ideals, advantages, and disadvantages. The survey offers an overview of T2V generation techniques but lacks a comprehensive exploration of datasets and evaluation metrics.

Xing et al.[14] provide a detailed overview of T2V generation methods, including datasets and evaluation metrics. However, this survey is somewhat outdated and focuses primarily on diffusion model-based architectures. In contrast, our survey covers all types of architectures for T2V generation except diffusion models.

Cho et al.[13] provides an excellent introduction for beginners, covering T2V applications, technical limitations, ethical conflicts, and future directions. However, their work has limitations, including an insufficient introduction to mainstream methods, insufficient comprehensive coverage of datasets, especially the lack of introduction to newly proposed datasets[10, 15], and lack of introduction to metrics such as EvalCrafter[16] and FETV[17].

In contrast, our survey not only comprehensively introduces related research, including core ideas, strengths, and weaknesses, but also introduces T2V datasets and evaluation metrics in detail. We cover 40 datasets and 20 metrics, overcoming the limitations of existing surveys.

III. Text-to-Video Generation

1. Preliminaries

The primary generation procedure is illustrated in Figure 3. First, a text encoder processes the input text to encoded features. These features are then utilized to produce the corresponding video by a generative model.

Text Encoder. Existing text encoders can be divided into two categories: pre-trained multimodal models such as

Chinese Journal of Electronics

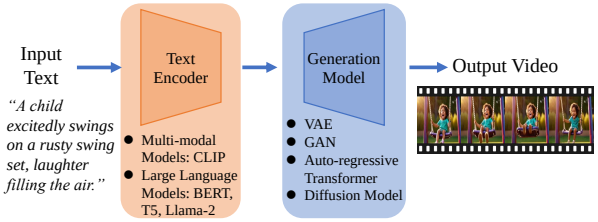


Figure 3 A brief diagram of text-to-video generation.

CLIP[18], and pre-trained large language models (LLMs) such as BERT[19], T5[20], and Llama-2[21].

Pre-trained multimodal models, exemplified by CLIP[18], learn their matching relationship by training on large-scale text-image pairs, thereby aligning image and text in an embedding space. However, CLIP cannot handle text with complex meanings, which may limit its effectiveness for long and complex text input.

Pre-trained LLMs excel in various tasks after being trained on large-scale corpora. BERT[19] can learn from unlabeled data and exhibit impressive performance, which can be further improved as model size and training data expand. T5[20] and Llama-2[21] are favored for their superior performance and the availability of open-source. Usually, LLMs outperform CLIP in understanding long text input.

Generation Model. Existing methods for generations can be divided into four categories: 1) VAE-based[22] approaches, 2) GAN-based[23] approaches, 3) Autoregressive transformer-based[4] approaches, 4) Diffusion model-based[24] approaches, and 5) T2I methods for video generation methods. Figure 4 shows the timeline of representative T2V generation methods in academia and industry. Figure 5 shows the categorization of existing methods for T2V generation.

2. VAE-based Approaches

The Variational Autoencoder (VAE)[22] is a groundbreaking method for generating images. It consists of an encoder and a decoder. The encoder maps the input data into a probability distribution, while the decoder generates new data by sampling from the learned probability distribution. Sync-DRAW[25] and GODIVA[26] are representative T2V generation methods based on VAE.

Sync-DRAW[25] combines a VAE with a recurrent attention mechanism for generating videos. It generates temporally coherent video frames by focusing on individual frames through the attention mechanism, while using the VAE to globally learn the latent distribution of the video. In addition, it keeps full attention to the object through a gating mechanism, which can generate videos that maintain the structural integrity of the object.

The GODIVA model[26] is the first to use VQ-VAE[107] for open-domain T2V generation, as illustrated in Figure 6. It combines VQ-VAE and 3D sparse attention to generate video, where 3D sparse attention can significantly reduce the computational cost. First, a VQ-VAE autoencoder is trained to represent continuous video pixels as discrete tokens. Then, a 3D sparse attention model is trained using language as input, with the discrete video tokens used as labels for video generation.

3. GAN-based Approaches

Generative Adversarial Networks (GANs)[23] have been ruling image generation for a decade. In contrast to VAE, the core idea of GAN is to estimate the generator via an adversarial process. GANs can usually produce images with good perceptual quality and are widely used in T2V generation methods. However, the models based on GAN can only generate videos with moving digits or simple human actions. They cannot further generate more complex and diverse videos anymore. Moreover, GAN often suffers from pattern collapse problems, and it is also difficult to scale these methods to complex and large-scale video datasets.

TGANs-c[27] is the first GAN-based work for T2V generation, proposing a temporal GAN where the generator takes text embeddings and noise vectors to produce video frames. It enhances the traditional 2D generator to a 3D model for better spatiotemporal dynamics and incorporates motion analysis in the discriminator to ensure coherent frame transitions, as detailed in Figure 7. TFGAN[28] introduces a new text conditioning method for discriminator feature maps through convolutional operations, as depicted in Figure 8. Meanwhile, the authors also created the Moving Shapes Dataset, where the text describes the shapes moving along a trajectory. VGFT[29] proposed a hybrid generator that combines GAN with VAE[22] to extract statistical and dynamic information from text, thereby generating diverse and smooth videos that correspond well to the input text.

Leverage the previous frame for generation. IRC-GAN[30] proposes the introspective recurrent convolutional GAN, consisting of the Recurrent Transconvolutional Generator (RTG) and Mutual-information Introspection (MI). RTG generates each frame based on the previous one to obtain better coherence. MI uses mutual information to compute the semantic distance between the T2V and the corresponding text and tries to minimize the distance.

TiVGAN[31] proposes a new training framework that initially focuses on learning the relationship between text and image to create high-quality single video frames. As training progresses, the model is gradually trained on more successive frames, which can stabilize the training and allow for clearer video generation.

Story visualization. StoryGAN[32] proposes a new task called "story visualization." The input is a multi-sentence paragraph, a story, and the output is a series of visualization images, with one for each sentence. Compared to other T2V works, this task can focus less on the continuity of the gen-

Chinese Journal of Electronics

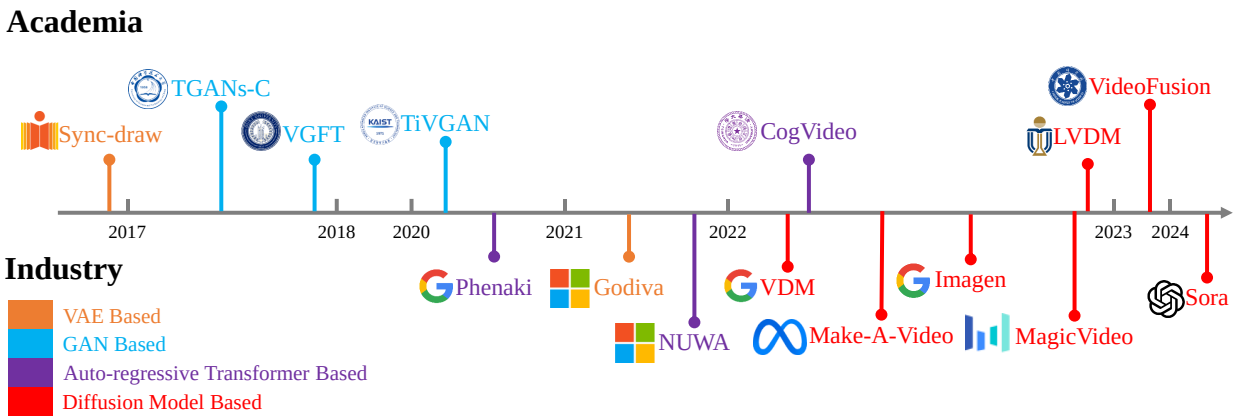


Figure 4 A timeline of representative text-to-video generation methods in academia and industry.

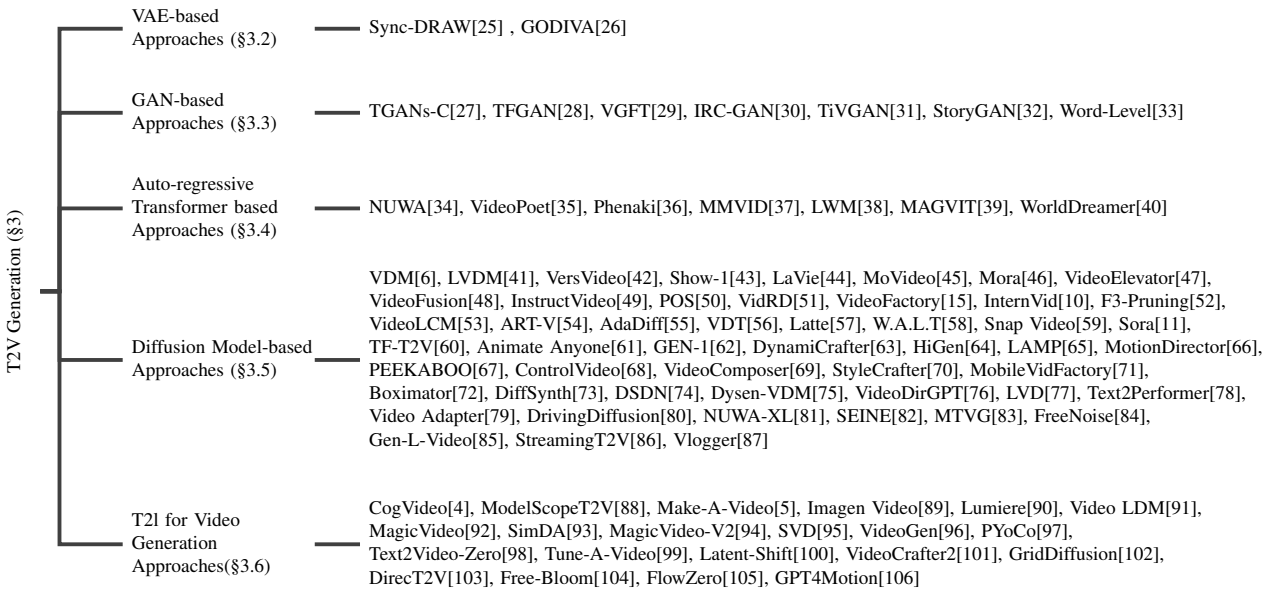


Figure 5 The categorization of existing methods for text-to-video generation.

erated image frames and more on the global consistency between dynamic scenes and characters.

Word-Level[33] expands on StoryGAN[32] by introducing a new sentence representation that combines word information from all story sentences. Also, a new fusion feature discriminator is proposed, extending spatial attention to improve image quality and story consistency.

4. Auto-regressive Transformer Based Approaches

The autoregressive approach allows the model to generate sequences step-by-step, with the latest generated content being based on the previously generated content, which naturally fits with the idea of generating coherent videos. Compared with GAN-based methods, the autoregressive transformer-based can avoid pattern collapse problems and generate better video quality. However, these methods require more computational and memory resources as the intermediate processes need to be stored and are constantly involved in the computation.

NUWA[34] introduces a 3D transformer encoder and decoder framework that provides a unified representation space for images and video, supporting both T2I and T2V generation. A 3D Nearby Attention (3DNA) mechanism is proposed to reduce the computational complexity. The architecture is shown in Figure 9.

VideoPoet[35] utilizes the unified architecture of LLMs to perform unified autoregressive learning for text, video, image, and audio modalities. Each modality has a separate tokenizer that converts the data into a discrete sequence of to-

Chinese Journal of Electronics

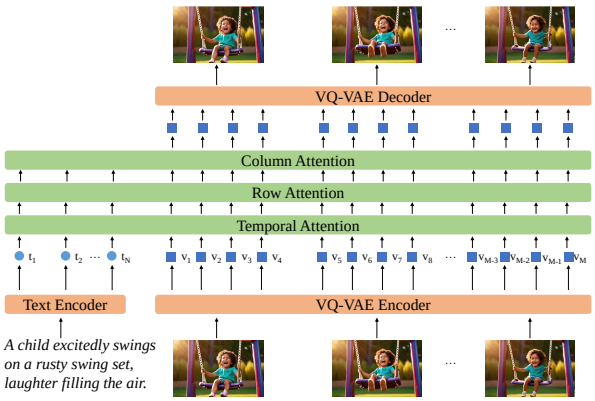


Figure 6 The architecture of GODIVA[26].

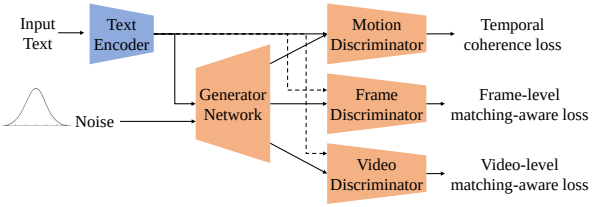


Figure 7 The architecture of TGANs-C[27].

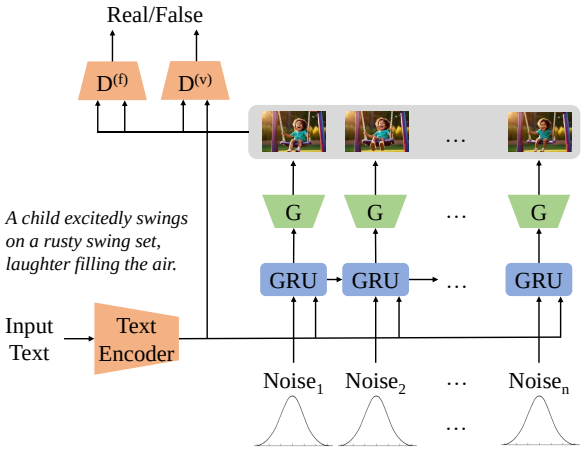


Figure 8 The architecture of TFGAN[28].

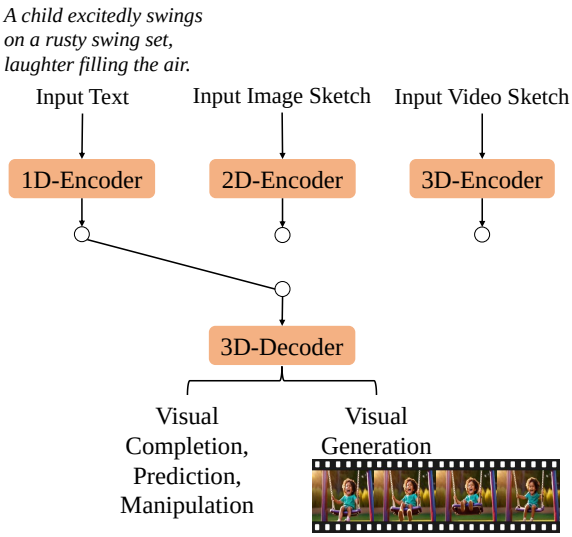


Figure 9 The architecture of Nuwa[34].

kens. In addition, it incorporates a super-resolution module in the token space to improve video quality.

Generate variable-length videos. Phenaki[36] is the first model to generate videos of variable length. First, it trains the transformer and randomly masks the video tokens during training. While generating the video, an arbitrary length video is generated by freezing the past tokens.

MMVID[37] is a video generation model that accepts multimodal inputs. It consists of an auto-encoder for encoding images and videos and a non-autoregressive transformer for predicting video tokens from multimodal inputs, which can generate better temporally consistent videos, such as special VID token, textual embedding, and improved mask prediction. Moreover, it proposes a new dataset called Multimodal VoxCeleb, where the video sources are VoxCeleb's[108] 19,522 videos, and each video has 36 manually labeled facial attributes.

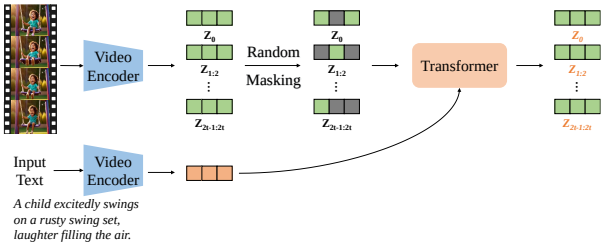
Current language models fall short of understanding the world, which is not easily described in words, and need help handling complex, long-form tasks. LWM[38] attempts to address this problem. It proposes the RingAttention technique, which extends the context window of the model so that it can handle sequences up to one million tokens long. The context length is gradually increased during the training process, and the training data includes text-image pairs, text-video pairs, and chat data for downstream tasks.

Some methods are not based on autoregressive transformers, but their architectures are based on transformers, such as MAGVIT[39] and WorldDreamer[40]. MAGVIT can handle multiple video synthesis tasks simultaneously and significantly outperforms contemporaneous diffusion and autoregressive methods in inference speed. WorldDreamer is the first generalized world model built for video generation. It proposes the spatial temporal patchwise Transformer (SPT), which performs attentional manipulation of local patches within a spatio-temporal window. SPT facilitates the learning of visual signal dynamics and accelerates the convergence of the training process, leading to about three times faster than diffusion-based methods.

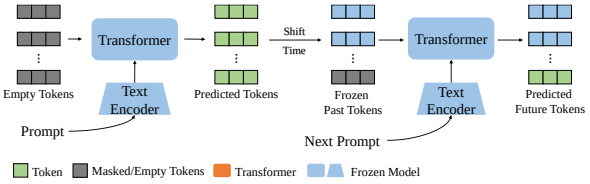
5. Diffusion Model-based Approaches

Denoising diffusion probabilistic model[24] (DDPM) can avoid the problem of pattern collapse in GANs and low generation quality in VAEs. At its core, a diffusion model adds random noise to existing data and reverses the process to generate the high-quality image. Through this process, the model

Chinese Journal of Electronics



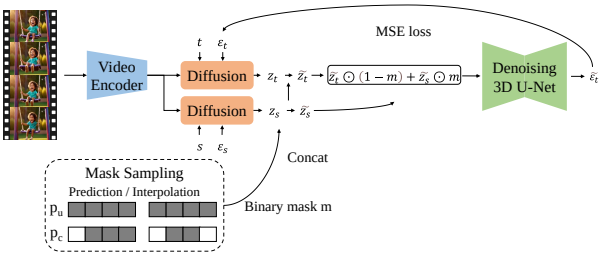
(a) Transformer training process.



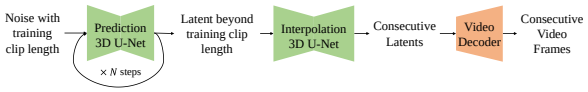
(b) Video generation process.

Figure 10 The architecture of Phenaki[36].

guidance, which effectively mitigates the cumulative error while extending to have more than one thousand frames.



(a) Training process.



(b) Inference process.

Figure 11 The architecture of LVDM[41].

learns to create synthetic data.

Since the diffusion model manipulates images in pixels, the computational consumption is particularly massive for large images. Rombach R et al.[109] proposed an image generation model based on the Latent Diffusion Model, whose core idea is to utilize an encoder an image to a latent vector and a decoder to decode the latent vector into an image. The advantage of such an approach is that the operation of the diffusion model is performed in the latent space, whose dimension is much smaller than the original pixel space. Thus, the computational consumption is significantly reduced.

The work introduced below is all based on either the diffusion model or the potential diffusion model, and no specific distinction is made in this survey.

VDM[6] is the first work to employ diffusion model for video generation. It extends the traditional U-Net[110] architecture (2D) to 3D spatiotemporal and supports joint training of images and videos at the same time. It further proposes conditional sampling for spatio-temporal video, which is capable of generating long and high-resolution videos. With the introduction of the 3D U-Net architecture, there has been an increase in the use of diffusion modeling for video generation.

Temporal modeling exploration. LVDM[41] is one of the representative works of the latent diffusion model in video generation. It innovatively proposes hierarchical diffusion in the latent space. The framework is shown in Figure 11, where t and s are randomly sampled diffusion timesteps for generated latents and conditional latents, respectively. p_c and p_u are probabilities of the conditional and unconditional input, respectively. After that, to overcome the performance degradation problem caused by long video generation, LVDM[41] proposes conditional latent perturbation and unconditional

Arguing that the oversimplification of other works for temporal modeling limits the spatiotemporal performance, VersVideo[42] proposes multi-excitation paths for spatiotemporal convolution across a pool of dimensions with different axes and multi-expert spatiotemporal attention blocks, which improves the spatiotemporal performance of the model without significantly increased training and inference costs. It also integrates the temporal module into the decoder to solve the problem of information loss due to the latent space.

Multi-stage T2V methods. Show-1[43]innovatively combines the respective advantages of pixel-based and latent-based video diffusion models. Pixel-based video diffusion models perform well but have high computational costs. While the latent-based model effectively reduces the computational effort, it is not easy to accurately align text and video. Show-1 first generates a low-resolution video using the pixel-based diffusion model to accurately align the text and video. Then, the latent video diffusion model is used to upsample the low-resolution video to high-resolution video. Its framework is shown in Figure 12.

LaVie[44] consists of three modules: a basic T2V model, a temporal interpolation (TI) model, and a video super-resolution (VSR) model. The basic model generates keyframes, the TI model generates smoother results, and the VSR model further improves resolution. MoVideo[45] generates the video in two steps, first generating the depth and optical flow of the video and then generating the final video by combining the keyframes generated by the T2I model under these two conditions. Mora[46] utilizes a variety of advanced large models for T2V generation that can replicate Sora's[11] generative capabilities. Specifically, video generation is decomposed into several subtasks, each assigned to a specialized large-scale model. VideoElevator[47] uses encapsulated T2V models to improve temporal consistency and

Chinese Journal of Electronics

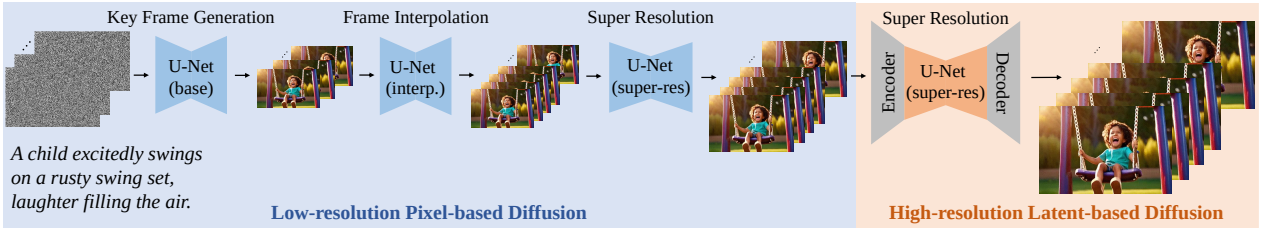


Figure 12 The architecture of Show-1[43].

T2I models to provide high-quality detail.

Multi-stage methods improve the quality of generated videos better than single-stage methods. However, the drawbacks are the complexity of the generation process and the increased burden of training.

Noise prior exploration. VideoFusion[48] decomposes the standard diffusion process into adding basic and residual noise, where consecutive frames share the basic noise. This way, frames in the same video clip are encoded as related noises, allowing the denoising network to reconstruct coherent video more easily. InstructVideo[49] combines human preferences with text into noise. POS[50] proposes optimal noise approximators and semantic-preserving rewriters. The optimal noise approximator first searches for the video closely related to the text and then inverts it into the noise space as an improved noise for the text input. The semantic preservation rewriter rewrites the original text while preserving the semantics.

The generated video is denoised from noise, which can directly affect the video quality. Improving the initial noise without changing other modules can further improve the model's performance.

Datasets contribution. VidRD[51] proposes a set of strategies for combining video-text data that involves different elements of several existing datasets, including video datasets for action recognition and image-text datasets. VideoFactory[15] collected videos on YouTube and labeled them using BLIP-2, building a large video dataset called HD-VG-130M. InternVid[10] proposes a method for autonomously building high-quality video text datasets using LLMs and publicly releasing the collected datasets.

High-quality T2V datasets are essential for improving model performance. Compared to other generative tasks, the available datasets in T2V generation are currently quite scarce, which limits the model performance to some extent.

Efficient training. F3-Pruning[52] proposes a training-free generalized pruning strategy to prune redundant spatiotemporal attention weights. This speeds up the inference of the T2V model and ensures video quality. VideoLCM[53] applies consistency models (CM)[111] to the video generation domain. It achieves high-fidelity and smooth video synthesis using only four sampling steps, demonstrating the potential of real-time synthesis. ART-V[54] learns only simple continuous motions between neighboring frames and generates videos autoregressively, thus reducing the enormous computational overhead of training. AdaDiff[55] arranges denoising steps according to different samples. It uses the gradient method for optimization to maximize a well-designed reward function. It reduces the inference time by at least one-third while achieving similar results to other methods.

Incorporating transformers into diffusion models. VDT[56] is the pioneer in using transformers in diffusion-based video generation. Utilizing transformer in diffusion models can leverage its rich spatiotemporal representations well. Similar to VDT, Latte[57] is also a transformer-based diffusion model that achieves a performance beyond that of the VDT. W.A.L.T[58] devised a transformer framework that transforms the latent vectors of images and videos in the same latent space. The transformer framework is a window attention architecture consisting of self-attention layers that alternate between non-overlapping, window-constrained spatial and spatio-temporal attention. Snap Video[59] replaces U-Nets in the traditional diffusion model with a transformer-based FITs[112] structure. Further, it extends the number of parameters, significantly improving temporal consistency and motion modeling.

Sora[11] is the first large-scale general-purpose video generation model that has attracted widespread attention in the community. It is based on a DiT[113] structure similar to Latte. It has several features. The first is the ability to train based on videos and images with different resolutions, durations, and aspect ratios and trained on the original image size. Secondly, it will convert short user prompts into longer detailed instructions using GPT[114] to improve video quality. Thirdly, it supports generation from images and videos, including image-to-video generation, extended generated video, video editing, etc.

Multiple data training. Considering the limited scale of publicly available text-video pairs, TF-T2V[60] proposes a new T2V generation framework that allows direct learning using text-free videos. The basic principle is to separate the process of text decoding from the process of temporal modeling. For this purpose, it employs a content and a motion branch, jointly optimized with shared weights. In the content branch, paired image-text data is leveraged to learn text-conditioned and image-conditioned spatial appearance gen-

Chinese Journal of Electronics

eration. The motion branch supports the training of motion dynamic synthesis by feeding text-free videos (or partially paired video-text data if available).

Personalized video generation. Animate Anyone[61] proposes a new framework customized for character animation, capable of converting character photos into animated videos controlled by a required sequence of poses while ensuring a consistent appearance and temporal stability. GEN-1[62] proposes a structure- and content-oriented video diffusion model that can modify existing videos based on text. DynamiCrafter[63] animates open-domain images using a pre-trained video diffusion prior, a proposed dual-stream image injection mechanism, and a dedicated training paradigm.

Spatio-temporal decoupling. HiGen[64] improves performance by decoupling the spatial and temporal elements of video from both structure and content perspectives. At the structural level, it uses a unified noise reducer to decompose the T2V task into two steps: spatial inference and temporal inference. At the content level, it extracts motion and appearance changes from the content of the input video, respectively. LAMP[65] proposes a new setting for the T2V generation task to balance the generation freedom with the training cost. It learns the motion patterns only from a training set consisting of 8 to 16 videos and later generates subsequent frames using the images generated by the T2I model as the first frame. MotionDirector[66] utilizes a dual-path LoRAs architecture to decouple the learning of appearance and motion and designs a new appearance debiasing temporal loss to mitigate the effect of appearance on the temporal training objective.

Controllable T2V generation. PEEKABOO[67] expects users to control video generation interactively. It proposes a new spatial-temporal masked attention module to achieve spatiotemporal control without the extra overhead of training and inference for current video generation models. ControlVideo[68] is adapted from ControlNet[115] and introduces three new modules to improve video generation. First, full cross-frame interaction is added to the self-attention module. Second, it uses frame interpolation to mitigate flicker effects. Finally, it synthesizes multiple consistent short clips. VideoComposer[69] offers a variety of methods for controllable video generation at once. It can simultaneously control spatial and temporal patterns in composited video through textual descriptions, sketch sequences, reference videos, and even simple manual movements. StyleCrafter[70] augments the pre-trained T2V model with a style control adapter, which can generate videos in any style by providing reference images. It uses a style-rich image dataset to train the style control adapter. MobileVidFactory[71] is a system that uses text input to generate videos for mobile devices automatically. The system first generates high-quality videos using a video generator. Then, the user can enrich the visual presentation by adding specified text. Finally, it matches the generated video with the appropriate audio in an audio database. Boximator[72] is a method for fine-grained video motion control that provides two types of boxes, allowing the user to select any object and define its motion without entering additional text. It trains only the control module, which retains knowledge of the underlying model, so its performance improves as the underlying model evolves.

Remove flicker and artifacts. Flickering and artifacts in the generated video are due to the current model's lack of learning and generative capabilities. Removing the flickering and artifacts can make the generated video more realistic. DiffSynth[73] proposes a latent in-iteration deflickering framework and a video deflickering algorithm to mitigate the flickering. The latent in-iteration deflickering framework applies the video deflickering algorithm to the latent space of the diffusion model, effectively preventing the accumulation of flicker in intermediate steps. The video deflickering algorithm remaps objects in different frames and blends them to enhance video consistency. Like DiffSynth[73], DSDN[74] also reduces flicker and artifacts in the generated video. It designs two diffusion streams, one for the video content and one for the motion variations, in such a way that the content and the motion can be better aligned. Experiments show that this decomposition also reduces the generation of flicker.

Complex dynamics modeling. Dysen-VDM[75] proposes a dynamic scene manager module to enhance the dynamics of generated videos. The module consists of 1) extracting key actions from the input text in chronological order, 2) converting the action schedule into a dynamic scene graph representation, and 3) enriching the scenes in the DSG with sufficiently reasonable details. VideoDirGPT[76] inputs the text prompts into GPT-4[114] to output a video plan, which includes generating scene descriptions, entities with their respective layouts, backgrounds for each scene, and consistent grouping of entities and backgrounds. Finally, the video generator generates the video based on the video plan. LVD[77] utilizes LLMs to generate dynamic scene layouts based on the prompt and then uses the generated layouts to guide the diffusion model to generate video. Such a process does not involve any updates to the parameters of the LLM and the diffusion model.

All three methods leverage the comprehension capabilities of large language models to guide generative models to better generation.

Domain-specific T2V generation. Text2Performer[78] focuses on the generation of human videos. It has two novel designs: decomposed human representations and a diffusion-based motion sampler. Video Adapter[79] decomposes domain-specific video distributions into pre-trained priors and trainable components, which significantly reduces the cost of tuning large pre-trained video models. DrivingDiffusion[80] generates realistic multi-view driving videos from prompts and 3D layouts.

Chinese Journal of Electronics

Generating longer videos. NUWA-XL[81] is a follow-up work of NUWA[34] generating long videos from text. It employs a coarse-to-fine generation paradigm. A global diffusion model generates keyframes over the entire period, and then a local diffusion model recursively fills in the content between nearby frames. SEINE[82] proposes a short-to-long (S2L) video diffusion model. It generates transitions based on textual descriptions automatically. Transition videos are generated by providing images of different scenes as inputs, combined with text-based control. MTVG[83] proposes multi-T2V generation that directly utilizes a pre-trained diffusion-based T2V generation model without additional fine-tuning. Similarly, FreeNoise[84], like MTVG[83], studies the generation of long videos conditioned on multiple texts. Instead of initializing noise for all frames, FreeNoise rearranges a series of noises for long-range correlation and provides temporal attention to them through window-based fusion. Gen-L-Video[85] extends existing short video diffusion models to generate long videos based on hundreds of clips with different semantics without introducing additional training while maintaining content consistency. StreamingT2V[86] proposes an autoregressive approach that utilizes novel short-term and long-term dependency blocks to seamlessly carry over video chunks with high motion while preserving high-level scene and object features during the generation process. Vlogger[87] is a system that generates vlogs longer than 5 minutes from text. It utilizes the LLM as a director and breaks down the generation of vlogs into four phases: Script, Actor, showmaker, and Voicer.

6. T2I for Video Generation Approaches

Training a T2V model from scratch requires tremendous computational cost. Thus, many works focus on how pre-trained T2I models can be utilized to contribute to video generation. The T2I-based model reduces the training cost and ensures the image quality of the generated video.

CogVideo[4] generates several key frames using a T2I model called CogView2[3]. Based on the keyframes, several rounds of frame interpolation are performed to form a final video. The process of frame interpolation is an autoregressive process. It also proposes multi-frame rate hierarchical training to align text-video pairs better. The framework is shown in Figure 13.

ModelScopeT2V[88] is the first open-source diffusion-based T2V generation model. Spatio-temporal blocks are added to a T2I synthesis model to ensure consistent frame generation and smooth motion transitions.

Make-A-Video[5] incorporates a super-resolution module based on frame interpolation to improve video quality. There is no need for text-video pairs for its training, and only video data is needed to learn the motion.

Imagen Video[89] utilizes the mature T2I model Imagen[116] to generate the base video. Six diffusion models are then cascaded, three for spatial super-resolution and three

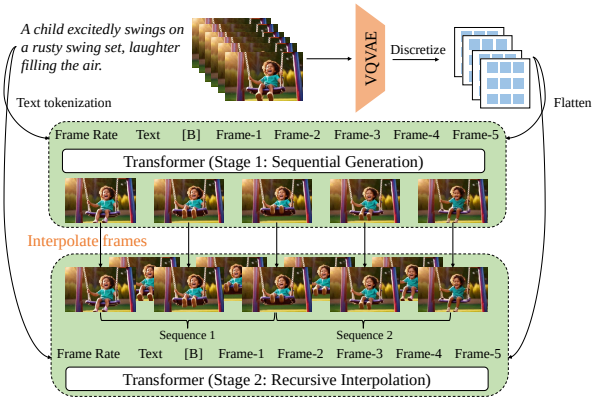


Figure 13 The architecture of CogVideo[4].

for temporal super-resolution. Each model is trained independently, and the cascade maximizes the performance benefits. The framework is shown in Figure 14.

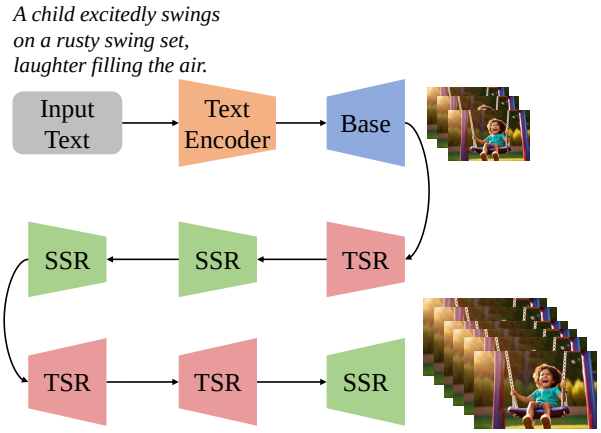


Figure 14 The architecture of Imagen[89].

Unlike generating keyframes and then interpolating them, Lumiere's[90] proposed STUNet can directly generate all the frames in one step and then use spatial super-resolution on some overlapping windows to get higher-resolution video.

Video LDM[91] first pre-trains the image generator on images. It then introduces a temporal layer and fine-tunes the encoded image sequence to convert the image generator into a video generator.

While previous approaches usually added a 1D temporal layer to model the time, MagicVideo[92] considered it unnecessary to use such a complex operation and proposed the concept of the frame adapter, which uses only two sets of parameters to model the relationship between the images and the video.

Similar to MagicVideo[92], SimDA[93] employs adapters to transform T2I models into T2V models. It not only includes a lightweight spatial adapter to transfer visual information for T2V learning but also introduces a temporal

Chinese Journal of Electronics

adapter to model temporal relationships for lower feature dimensions.

MagicVideo-V2[94] integrates a T2I model, a video motion generator, a reference image embedding module, and a frame interpolation module into an end-to-end video generation pipeline.

SVD[95] proposes a three-step paradigm for training video generation models: T2I pre-training, video pre-training, and high-quality video fine-tuning. In addition, it provides a series of processes to generate high-quality T2V datasets.

VideoGen[96] utilizes the T2I model to generate a reference image based on the prompt. Then, an efficient cascading latent diffusion model is introduced, which conditions the reference image and prompt for generating the latent representation of the video.

PYoCo[97] proposes a video diffusion noise for fine-tuning T2I models into T2V models. It fine-tunes the eDiff-I[117] to construct a large-scale T2V diffusion model.

Text2Video-Zero[98] utilizes a pre-trained T2I model to generate the latent space representation of the image. After that, the latent space representation of each frame is generated using the dynamics method and the cross-attention mechanism that only pays attention to the first frame. Finally, the video is generated by the decoder.

Tune-A-Video[99] proposes a new task of training a T2V model using only a single text-video pair and a pre-trained T2I model.

Latent-Shift[100] proposes a parameter-free temporal shift module that can generate videos based on the T2I model. The module accomplishes this by shifting both parts of the feature mapping channel forward and backward along the time dimension.

VideoCrafter2[101] separates appearance and motion by utilizing low-quality videos for motion learning and high-quality images for appearance learning. It also suggests using synthetic images with complex concepts instead of real images for fine-tuning.

GridDiffusion[102] generates videos using the grid diffusion model. It first generates key grid images, including four images inside a grid image. After that, masked grid images are inserted into the grid, allowing the interpolation model to generate the masked images autoregressively.

DirecT2V[103] and Free-Bloom[104] use language models to transform user prompts into detailed frame descriptions, then employ a T2I model to generate each frame. DirecT2V enhances frame consistency using novel value mapping and dual softmax filtering, while FreeBloom proposes joint noise sampling and dual path interpolation. FlowZero[105] utilizes LLM to generate a comprehensive dynamic scene syntax (DSS) containing scene descriptions, object layouts, and background motion patterns. The DSS then guides the image diffusion model in generating videos. In particular, it proposes a self-refining iterative process that enhances the alignment of the video with the text. GPT4Motion[106] utilizes GPT-4[114] to generate Blender scripts based on user prompts, producing coherent physical motion across frames. Blender* is an open-source 3D creation suite that provides modeling, animation, and rendering tools that facilitate the creation of detailed 3D scenes.

7. Open-source Organization for T2V Methods

Compared to other research fields, T2V requires a lot of computational and data resources, and the models are usually released by industry. For commercial reasons, many models and training details are not open source. We summarize the existing open-source methods in Table 2 to help researchers quickly get started with the experiments.

IV. Datasets

The dataset for the T2V task can be categorized into two classes based on the text[14]. The first is caption-level datasets, where the text corresponding to the video is more detailed in description, and the other is category-level datasets, where the text corresponding to the video is a category of the video.

1. Caption-level Datasets

We list the current common caption-level datasets in the T2V task in Table 3. From the table, we can observe that the early datasets annotated with text are manually annotated (Manual), and the videos are small in size and single domain (e.g., movie, action, cooking), as well as low resolution (e.g., 240P). With the release of WebVid-10M[122], the T2V dataset has ushered in an era of rapid development, and it has become the most dominant dataset in the T2V task. However, the resolution of WebVid-10M[122] is too low, and a watermark exists, leading to poor video quality. Therefore, subsequent datasets have increased the video resolution and added algorithms to filter inappropriate videos (e.g., the presence of watermarks or subtitles).

In addition to gradually improving the quality of the videos in the dataset, the newly released datasets also pay more attention to the alignment between text and video. Improving the alignment between text and video improves the generation performance of the model, which has been demonstrated in recent work[10, 95].

Manual annotation can provide high-quality text, but if the number of videos rises, the burden of manual labor will be unbearable. HowTo100M[123] and other datasets collect videos originating from YouTube, and they use the automatic speech recognition (ASR) technique provided by YouTube to generate the texts, but the semantic relevance is low. WebVid10M[122] uses Alt-text, and WTS70M[124] uses Metadata (which contains titles, descriptions, tags, and channel names). VideoCC3M[125] transfers the text-image dataset to the text-video dataset. It uses Conceptual

*https://www.blender.org/

Table 2 Open source T2V methods collation.

Method	Venue	Frames	Resolution	Code	Official Release
Follow Your Pose[118]	AAAI24	8	512 × 512	https://github.com/mayuelala/FollowYourPose	✓
ConditionVideo[119]	AAAI24	24	512 × 512	https://github.com/pengbo807/ConditionVideo	✓
Make-A-Video[5]	Arxiv22	16	256 × 256	https://github.com/lucidrains/make-a-video-pytorch	×
LVDM[41]	Arxiv22	16	256 × 256	https://github.com/YingqingHe/LVDM	✓
DirecT2V[103]	Arxiv23	16	512 × 512	https://github.com/KU-CVLAB/DirecT2V	✓
LaVie[44]	Arxiv23	61	$1280x2048$	https://github.com/Vchitect/LaVie	✓
ModelScope[88]	Arxiv23	16	256 × 256	https://modelscope.cn/models/iic/text-to-video-synthesis/summary	✓
VidRD[51]	Arxiv23	16	256 × 256	https://github.com/anonymous0x233/ReuseAndDiffuse	✓
VideoDirectorGPT[76]	Arxiv23	16	256 × 256	https://github.com/HL-hanlin/VideoDirectorGPT	✓
Show-1[43]	Arxiv23	29	320 × 576	https://github.com/showlab/Show-1	✓
VideoFusion[48]	Arxiv23	33	256 × 256	https://github.com/ai-forever/KandinskyVideo	✓
HiGen[64]	Arxiv23	32	448 × 256	https://github.com/ali-vilab/VGen	✓
Animate Anyone[61]	Arxiv23	24	768 × 768	https://github.com/HumanAIGC/AnimateAnyone	✓
StyleCrafter[70]	Arxiv23	16	320 × 512	https://github.com/GongyeLiu/StyleCrafter	✓
DynamiCrafter[63]	Arxiv23	16	576 × 1024	https://github.com/Doubiiu/DynamiCrafter	✓
MotionDirector[66]	Arxiv23	16	384 × 384	https://github.com/showlab/MotionDirector	✓
FlowZero[105]	Arxiv23	8	512 × 512	https://github.com/aniki-ly/FlowZero	✓
Latte[57]	Arxiv24	16	256 × 256	https://github.com/Vchitect/Latte	✓
VideoCrafter2[101]	Arxiv24	16	320 × 512	https://github.com/AILab-CVC/VideoCrafter	✓
MMVID[37]	CVPR22	8	128 × 128	https://github.com/snap-research/MMVID	✓
MAGVIT[39]	CVPR23	16	128 × 128	https://github.com/google-research/magvit	✓
Text2Performer[78]	CVPR23	20	512 × 256	https://github.com/yumingj/Text2Performer	✓
Dysen-VDM[75]	CVPR24	16	256 × 256	https://github.com/scofield7419/Dysen	✓
BIVDiff[120]	CVPR24	8	512 × 512	https://github.com/MCG-NJU/BIVDiff	✓
LAMP[65]	CVPR24	16	320 × 512	https://github.com/RQ-Wu/LAMP	✓
Tune-A-Video[99]	ICCV23	32	512 × 512	https://github.com/showlab/Tune-A-Video	✓
Text2Video-Zero[98]	ICCV23	8	512 × 512	https://github.com/Picsart-AI-Research/Text2Video-Zero	✓
CogVideo[4]	ICLR23	16	480 × 480	https://github.com/THUDM/CogVideo	✓
LVD[77]	ICLR24	16	512 × 512	https://github.com/TonyLianLong/LLM-groundedVideoDiffusion	✓
AnimateDiff[121]	ICLR24	16	256 × 256	https://github.com/guoyww/AnimateDiff	✓
FreeNoise[84]	ICLR24	64	1024 × 576	https://github.com/AILab-CVC/FreeNoise	✓
VDM[6]	NeurIPS22	16	64 × 64	https://github.com/lucidrains/video-diffusion-pytorch	×
Free-Bloom[104]	NeurIPS23	6	512 × 512	https://github.com/SooLab/Free-Bloom	✓

Captions3M[126] as the original dataset. It starts with the text image dataset and, for each text image pair in the dataset, finds frames in the video that are similar to the image and then extracts short video clips around the matching frames and corresponds the text to those clips.

The latest datasets all use different generative methods to get the texts, which saves labor and also ensures that the quality of the texts is high.

HD-VG-130M[15] first cuts the video using PySceneDetect[†]. After cutting, the content of each video contains only one scene. After that, select the middle frame of the video and use BLIP2[127] to generate a textual description. This description will be used to describe the video. InternVid[10] has two scales to generate text, coarse and fine, where the coarse scale is generated in the same way as HD-VG-130M[15]. At the fine scale, Tag2Text[128] is used to generate text descriptions for each frame of the video. These text descriptions are then synthesized into a comprehensive description using a pre-trained language model. CelebV-Text[39] utilizes a semi-automatic template-based text generation strategy. An algorithm automatically labels attributes that are easy to label, and attributes that are difficult to label are labeled manually. Afterward, following the template, the attributes are filled in to get the final description of the video. Vimeo25M[44] uses Videochat[129] to generate text automatically. Panda-70M[130] utilizes multiple models (including VideoLLaMA[131], VideoChat[129], VideoChat[129] Text, BLIP-2[127], and MiniGPT-4[132]) to generate texts. After that, it fine-tunes Unmasked Teacher (UMT)[133] to help select the best one of the texts. In order to minimize the computational requirements, it proposes a student model to extract knowledge from the teacher model. VidProM[134] collected 1.67 million T2V prompts from real users. Based on the prompts, 6.69 million videos were generated by Pika[‡], Text2Video-Zero[98], VideoCraft2[101], and ModelScope[88]. MiraData[§] uniformly sampled eight frames for each video and arranged them into a large 2x4 grid image. Then, a one-sentence caption is generated for each video using Panda-70M's[130] caption model. After that, the generated captions are fed into GPT-4V[135] as auxiliaries along with the large 2x4 image to output multi-dimensional cap-

[†]https://www.scenedetect.com/

[‡]https://pika.art/home

[§]https://mira-space.github.io/

tions in one dialog round efficiently.

As shown in Figure 15, we give examples of text-video pairs from MSVD, MSR-VTT, WebVid10M, and Panda70M to illustrate the development of the T2V dataset. We show four frames from the selected video uniformly over time. If there are multiple text annotations, we select two sentences from them for the demonstration. For comparison, we resize the videos to the same size. Both MSVD and MSR-VTT have multiple text annotations for the same video. MSVD may even have incorrect text annotations. The video from MSR-VTT contains multiple scenes, and the others are single scenes. From WebVid10M to Panda70M, we can see more precise text annotation.

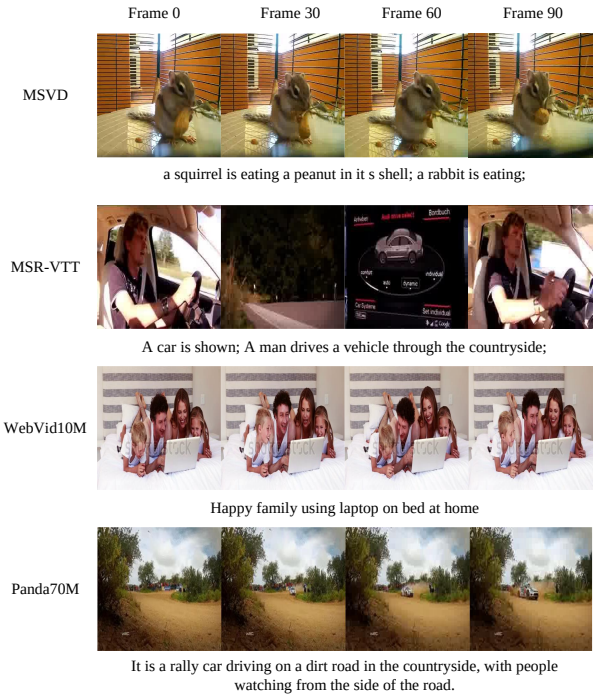


Figure 15 Showcase of different datasets.

2. Category-level Datasets

Without a suitable caption-level dataset, the T2V task uses category-level datasets to train the model. These category-level datasets are from other tasks, e.g., UCF101[150], Kinetics[151], and Something-Something[152] from the action recognition task. DAVIS[153] from the video editing task. We list the category-level datasets ever used for the T2V task in Table 4.

V. Evaluation Metrics

Quantitative metrics consist of the visual quality of T2V and the alignment of text and video. To better evaluate the performance of T2V models, EvalCrafter[16] further improves the metrics on visual quality and text-video alignment and proposes metrics on motion quality and temporal consistency. These will be introduced in the following four subsections. For qualitative metrics, which are subjective human evaluations, they will be introduced in Section 5.3.

Table 3 The comparison of main caption-level video datasets.

Dataset	Text	Domain	Clips	Res.
MSVD / 2011[136]	Manual	Open	2K	-
MSR-VTT / 2016[137]	Manual	Open	10K	240P
DideMo / 2017[138]	Manual	Flickr	27K	-
LSMDC / 2017[139]	Manual	Movie	118K	1080P
ActivityNet / 2017[140]	Manual	Action	100K	-
YouCook2 / 2018[141]	Manual	Cooking	14K	-
How2 / 2018[142]	Manual	Instruct	80K	-
VATEX / 2019[143]	Manual	Action	41K	240P
HowTo100M / 2019[123]	ASR	Instruct	136M	240P
WTS70M / 2020[124]	Metadata	Action	70M	-
YT-Temporal / 2021[144]	ASR	Open	180M	-
WebVid10M / 2021[122]	Alt-text	Open	10.7M	360P
Echo-Dynamic / 2021[145]	Manual	ECG	10K	-
Tiktok / 2021[146]	Mannual	Action	0.3K	-
HD-VILA / 2022[147]	ASR	Open	103M	720P
VideoCC3M / 2022[125]	Transfer	Open	10.3M	-
HD-VG-130M / 2023[15]	Generated	Open	130M	720P
InternVid / 2023[10]	Generated	Open	234M	720P
CelebV-Text / 2023[148]	Generated	Face	70K	480P
Vimeo25M / 2023[44]	Generated	Open	25M	-
Panda-70M / 2024[130]	Generated	Open	70M	720P
VidProM / 2024[134]	Collected	Open	6M	-
MiraData / 2024[149]	Generated	Game	57K	-

1. Video Quality Assessment

The traditional metrics to measure the visual quality of video are FVD[167] and IS[168], developed from image visual metrics.

Fréchet Video Distance (FVD)[167] builds on the principle of FID[169]. It measures the visual quality of the generated video by calculating the distance between the generated video's distribution and the real video's distribution. The cal-

Table 4 The comparison of main Category-level video datasets.

Datasets	Categories	Clips	Res.
KTH / 2004[154]	6	2K	160×120
MUG / 2010[155]	6	1K	896×896
UCF-101 / 2012[150]	101	13K	256×256
Cityscapes / 2015[156]	30	3K	256×256
Moving MNIST / 2016[157]	10	10K	64×64
Kinetics-400 / 2017[158]	400	260K	256×256
BAIR / 2017[159]	2	45K	64×64
DAVIS / 2017[153]	-	90	1280×720
Sky Time-Lapse / 2018[160]	1	38K	256×256
Ssthv2 / 2018[152]	174	220K	256×256
Kinetics-600 / 2018[161]	600	495K	256×256
MiT / 2018[162]	339	1M	340×256
Tai-Chi-HD / 2019[163]	1	3K	256×256
iPER / 2019[164]	10	206	256×256
Bridge Data / 2021[165]	10	7K	256×256
Mountain Bike / 2022[166]	1	1K	576×1024
RDS / 2023[91]	2	683K	512×1024

Chinese Journal of Electronics

culation formula is shown in Eq. 1,

$$d\left(P_R, P_G\right) = \left|\mu_R - \mu_G\right|^2 + \mathrm{Tr}\left(\Sigma_R + \Sigma_G - 2\left(\Sigma_R\Sigma_G\right)^{\frac{1}{2}}\right) \tag{1}$$

where μ_R and μ_G are the means, and $\sum_R$ and $\sum_G$ are the co-variance matrices of P_R and P_G, respectively. FVD[167] adopts inflated 3D Convnets[158] (I3D) pretrained on Kinetics[151] to extract features from videos.

Inception Score (IS)[168] uses the Inception Network[170], pre-trained on the ImageNet[171] dataset as the feature extraction to evaluate the image quality. When evaluating video quality, the feature extraction model is changed to 3D-Convnets (C3D)[172]. The calculation formula is shown in Eq. 2,

$$\mathrm{IS} = \exp E_{x\sim p_G} KL(p(y \mid x) \| p(y)) \tag{2}$$

where $P(y)$ is the marginal distribution of all videos and $P(y|x)$ denotes the output distribution of the model after inputting the generated videos. IS measures the diversity of the generated videos, with larger scores indicating more variety in the generated content.

A recent study, EvalCrafter[16], utilizes Dover[173] to assess the visual quality of generated videos, which consist of two components, VQA_A and VQA_T, which are the aesthetic and technical scores, respectively. The technical perspective involves quantifying the perception of distortions, while the aesthetic perspective focuses on preferences and recommendations about content.

2. Text-Video Alignment

In addition to video quality assessment, measuring the alignment between input text and generated video is another important perspective for evaluating T2V generation. The traditional evaluation metric is CLIPSIM[26], and EvalCrafter[16] further proposes more metrics to measure the text-video alignment more comprehensively. These evaluation metrics will be described below.

CLIPSIM[26] is calculated by first encoding the image and text with the CLIP[18] model to get the embeddings and then calculating the cosine similarity between the embeddings. The similarities between frames and the input text are averaged to represent the final similarity between the video and the input text. and then take the average value. The formula is described in Eq. 3,

$$CLIPSIM(p, x) = \frac{1}{t}\sum_{i=1}^{t} \mathcal{C}\left(\mathrm{emb}\left(x_t\right), \mathrm{emb}\left(p\right)\right) \tag{3}$$

where x_t^i means the $t-$th frame of the video, $emb(\cdot)$ means CLIP embedding, $\mathcal{C}(\cdot)$ means calculating the cosine similarity, and p means the text.

It is worth mentioning that the accuracy of CLIPSIM entirely depends on the CLIP[18] model. To reduce the side effect, **Relative Matching (RM)**[26] metric. CLIPSIM calculates the ratio of CLIPSIM of the generated video to that of the ground truth video. There are three other CLIPSIM-like metrics. **CLIPScore-ft** is based on the CLIP model fine-tuned on the MSR-VTT dataset[137]. **BLIPScore** and **UMTScore** use BLIP[127] and UMT[133] instead of CLIP.

In practical scenarios, limited by the performance of the CLIP model and the complexity of the prompt, the above traditional metrics can not work well. Therefore, a series of metrics are proposed in EvalCrafter.

SD-Score uses SDXL[174] to generate N_1 images per prompt, extracting the visual embeddings to calculate the similarity between the generated video and the SDXL images. Essentially, SDXL[174] acts as the teacher, and the video generation model as the student. The results generated by the student are close to those generated by the teacher. The calculation is shown in Eq. 4,

$$S_{SD} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N}\sum_{t=1}^{N}\left(\frac{1}{N_1}\sum_{k=1}^{N_1}\mathcal{C}\left(\mathrm{emb}\left(x_t^i\right), \mathrm{emb}\left(d_k^i\right)\right)\right)\right) \tag{4}$$

where x_t^i means the $t-$th frame of the $i-$th video. N_1 is typically set to 5.

BLIP-BLEU uses BLIP2[175] to generate the caption for the generated video and the BLEU[176] similarity between the caption and the prompt is calculated. Shown in Eq. 5,

$$S_{BB} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N_2}\sum_{k=1}^{N_2}\mathcal{B}\left(p^i, l_k^i\right)\right) \tag{5}$$

where $\mathcal{B}(\cdot,\cdot)$ is the BLEU similarity scoring function, $\left\{l_k^i\right\}_{k=1}^{N_2}$ are BLIP generated captions for $i-$th video, and N_2 is typically set to 5.

OCR-Score checks whether the text required to appear in the video appears in the generated video to test the model's ability to generate text. This process involves using PaddleOCR to detect the English text in the generated video, after that, calculate the word error rate (WER)[177], the normalized edit distance (NED)[178], and the character error rate (CER)[179]. The average of the three values is the OCR-Score.

Detection-Score detects whether the requested objects appear in the video,

$$S_{Det} = \frac{1}{M_1}\sum_{i=1}^{M_1}\left(\frac{1}{N}\sum_{t=1}^{N}\sigma_t^i\right) \tag{6}$$

where M_1 represents the count of prompts containing objects, and σ_j^i represents the detection result for frame t in video i (with a value of 1 indicating the detection of an object and 0 indicating otherwise).

Chinese Journal of Electronics

Count-Score detects whether the number of objects in the video is correct,

$$S_{\text{Count}} = \frac{1}{M_2}\sum_{i=1}^{M_2}\left(1-\frac{1}{N}\sum_{t=1}^{N}\frac{\left|c_t^i-\hat{c}^i\right|}{\hat{c}^i}\right) \tag{7}$$

where M_2 is the number of prompts with object counts, c_t^i is the detected object count frame t in video i and $\hat{c}^i$ is the ground truth object count for video i.

Color-Score detects whether the color in the video matches the description in the prompt,

$$S_{\text{Color}} = \frac{1}{M_3}\sum_{i=1}^{M_3}\left(\frac{1}{N}\sum_{t=1}^{N}s_t^i\right) \tag{8}$$

where M_3 is the number of prompts with object colors, s_t^i is the color accuracy result for frame i in video t (1 if the detected color matches the ground truth color, 0 otherwise).

Celebrity ID Score calculates the distance between the celebrity in the generated video and the real image of the celebrity,

$$S_{CIS} = \frac{1}{M_4}\sum_{i=1}^{M_4}\left(\frac{1}{N}\sum_{t=1}^{N}\left(\min_{k\in\{1,\ldots,N_3\}}\mathcal{D}\left(x_t^i,f_k^i\right)\right)\right) \tag{9}$$

where M_4 is the number of prompts that contain celebrities, $\mathcal{D}(\cdot,\cdot)$ is the Deepface's[180] distance function, $\left\{f_k^i\right\}_{k=1}^{N_3}$ are collected celebrities images for prompt i, and N_3 is set to 3.

3. User Study

Although, in the previous sections, we have introduced many automated evaluation metrics, some of these automated metrics have been found to be inconsistent with human judgments in some studies[3, 181, 182], indicating that automated evaluation metrics may not always be reliable. Therefore, the human perspective is also essential for evaluating generated videos.

There are four main benchmarks that are widely used by the public: DrawBench[183], FETV[17], EvalCrafter[16] and VBench[184].

DrawBench[183] is a benchmark for T2I generation, but it can also be used for T2V generation. The benchmark is proposed to compensate for COCO's[185] limited range of prompts, typified by the newly proposed PaintSkills[186], to systematically assess visual reasoning skills and social biases outside of COCO[185]. DrawBench has eleven evaluation categories with a total of 200 prompts. These categories include color, count, spatial positioning, conflicting interaction, long description, misspelling, rare words, quoted words, and so on.

FETV[17] is a fine-grained evaluation benchmark for T2V generation. It consists of 619 prompts, with 541 prompts sourced from existing datasets and 78 unique prompts created by the authors. Each prompt is categorized based on three aspects: the main content, attributes, and complexity. The feature referred to as "main content" was further divided into spatial and temporal categories. Similarly, "attribute control" encompasses both spatial and temporal qualities. The feature of "prompt complexity" is categorized into three levels: "simple," "medium," and "complex," which are determined by the number of consecutive words in the prompts. By employing classification, the FETV benchmark can be subdivided into distinct subsets, enabling fine-grained evaluation.

EvalCrafter[16] aims to create a list of reliable prompts to assess the capabilities of various T2V models fairly. To achieve its goal, EvalCrafter collected and analyzed a large number of prompts from the real world and selected more than 500. Afterward, EvalCrafter proposes an automated pipeline to increase the diversity of the selected prompts. In total, there are 50 styles and 20 camera motion prompts in the benchmark, and the average length of the prompts is 12.5 words, similar to real-world prompts.

VBench[184] is a comprehensive benchmark suite for video generative models. It decomposes video generation quality into 16 dimensions, and each evaluation dimension assesses one aspect of video generation quality. To reduce the overhead of generating videos, it accurately filters the set of tested prompts; for each metric, there are only 100 prompts. Experiments show that VBench's evaluation results align well with human perception.

Based on the benchmarks mentioned above, the researchers put the results generated by their model and those generated by others and asked the observer to choose the best-generated video based on certain aspects. The commonly examined aspects are video frame quality, semantic relevance, motion realism, etc.

In order to demonstrate the consistency of automatic assessment results with human assessment results, some studies[16, 17] calculate Spearman's rank correlation coefficients[187] and Kendall's rank correlation coefficients[188]. These coefficients reveal the direction and strength between automatic and human assessment scores.

4. Motion Quality Assessment

Previous T2V studies did not yet consider metrics for evaluating the motion quality of the generated video. EvalCrafter[16] propose Action-Score, Flow-Score, and Motion AC-Score for motion quality assessment.

Action Recognition (Action-Score) recognizes human actions in the generated video using the MMAction2 toolbox[189]. The action score is calculated as accuracy by comparing the recognized action with the action in the original prompt.

Average Flow (Flow-Score) uses the pre-trained optical flow estimation method RAFT[190] to extract the dense flow of the video in two-frame intervals. Then, calculate the aver-

Chinese Journal of Electronics

age flow score for the whole video clip. This helps identify static videos.

Amplitude Classification Score (Motion AC-Score). Based on the average flow, Motion AC-Score calculates the motion amplitude of the generated video and determines whether the amplitude is the same as the amplitude specified by the prompt. This gives us a clearer picture of the motion changes in the video.

5. Temporal Consistency Assessment

Warping Error. Firstly, using a pre-trained optical flow estimation network[190] to obtain the optical flow for every two frames, after that, the difference between the warped image and the predicted image is computed pixel by pixel, and the final score is the average of all pairs.

Semantic Consistency (CLIP-Temp). Specifically, calculates the semantic embedding on every two frames of the generated video, then obtains the average value of every two frames.

Face Consistency. This metric evaluates the human identity consistency of the generated video. It is calculated by selecting the first frame as the reference frame and calculating the cosine similarity between the embedding of the reference frame and the embeddings of other frames. The average of these similarities is taken as the final score.

VI. Experimental Results

Dataset. Currently, the T2V task is mainly evaluated in a zero-shot manner on the MSR-VTT[137] and UCF-101[150] datasets. MSR-VTT[137] consists of 10,000 video clips in 20 categories, each described by approximately 20 natural sentences. Typically, the textual descriptions corresponding to the 2,990 video clips in the test set were used as prompts to generate the corresponding videos. The UCF-101[150] consists of 13,320 video clips divided into 101 categories.

Evaluation Metrics. For the MSR-VTT[137] dataset, the FVD[167] and FID[169] metrics are used to evaluate the video quality, and CLIPSIM[26] is used to measure the alignment between text and video. For the UCF-101[150] dataset, the Inception Score, FVD[167], and FID[169] are used to evaluate the quality of the generated video and its frames. Many of the metrics mentioned are not yet widely used and are therefore not included in the statistics.

Comparison of Results. We summarize the experimental results of the most existing methods in Table 5. As illustrated, PixelDance[191] achieves the best FID on UCF-101, and VersVideo[42] achieves the best FVD and IS on UCF101. TF-T2V[60], ART•V[54], and MoVideo[45] achieved the best FID, FVD, and CLIPSIM on MSRVTT, respectively.

VII. Discussion

1. Challenges

Quantitative relationships in video. When a fixed number of objects is specified in the prompt, it is sometimes incorrectly reflected in the generated video. For example, the prompt mentions that two people are present, but the generated video has only one person throughout, or it changes from two people to some other number of people.

Causality of events. The model has difficulty understanding how actions and behaviors will drive events. An example would be coloring and painting a wall, but the wall color does not change.

Object interactions. The model has trouble understanding the boundaries between objects and modeling their interactions. For example, after throwing a ball, the ball and the basket merge instead of being bounced off.

Scale and proportionality. The model experiences difficulty understanding the relationship between scale size and proportion of different objects in different parts of the scene. For example, one person in the same scene is particularly short while another is particularly tall.

Object illusion. The objects generated by the model are unstable, appearing or disappearing suddenly in the video.

2. Future Trends

Large-scale open-source T2V datasets. Although many datasets have been proposed recently, the number is insufficient for the model to learn. Also, the quality of the videos and texts in the datasets needs to be continuously improved so that the model performance can be further enhanced. It is also essential to open source collected datasets, which can effectively accelerate the progress of the research.

Efficient training methods and model architecture. Training a T2V model takes a lot of computational effort and time. More efficient architectures reduce the time required for inference and weaken the hardware requirements needed for inference, which can significantly facilitate the application of the model.

Comprehensive metrics for evaluation. While the recent EvalCrafter[16] and FETV[17] have primarily filled the gap, the newly proposed metrics will be included in the methods comparison in the future.

Abstract text generation. Existing T2V generation methods all assume the input text is concrete, which is not always practical in the real world. For abstract words or abstract sentences, it is difficult for the model to generate well, and the quality of the generated video will drop significantly. For example, the prompt is “Hard work is a virtue.” However, such a demand is reasonable because people think abstractly, and abstract ideas can be challenging to describe. We hope that the results generated by the model can conform to our abstract thinking or help our abstract thinking to become more concrete.

Long video generation. Most of the research works mentioned in this survey can only generate short videos for 2 seconds with 16 frames, which limits the application use. If long videos with relatively high quality can be generated, T2V generation will have excellent application prospects.

Chinese Journal of Electronics

Table 5 Organization of experimental results on video generation methods.

Method	Venue	Training Dataset	Paired Data	Res.	MSRVTT[137]			UCF-101[150]		
					FID(↓)	FVD(↓)	CLIPSIM(↑)	FID(↓)	FVD(↓)	IS(↑)
LVD[77]	ICLR24	training free		512×512		521			861	
POS[50]	Arxiv23			256×256	42.29		0.2993		566.68	38.19
VDM[6]	NeurIPS22	UCF101[150]		64×64				298		57.62
NUWA[34]	ECCV22	VATEX[143]	241K	256×256	47.68		0.2439			
CogVideo[4]	ICLR23	WebVid[122]	5.4M	480×480	23.59	1294	0.2631	179	701.59	25.27
LVDM[41]	Arxiv22		2M	256×256		742	0.2381		641.8	
SimDA[93]	CVPR24		10M	256×256		456	0.2945			
VideoGen[96]	Arxiv23		10M	256×256			0.3127		554	71.61
ModelScope[88]	Arxiv23		10M	256×256	11.09	550	0.293		410	
Dysen-VDM[75]	CVPR24		10M	256×256	12.64		0.3204		325.42	35.57
VideoDirGPT[76]	2023		10M	256×256	12.22	550	0.2860			
PYoCo[97]	ICCV23		22.5M	256×256	9.73				355.19	47.76
VideoFusion[48]	CVPR23		10M	256×256		581	0.2795	75.77	639.9	17.49
Latent-shift[100]	Arxiv23		10M	256×256	15.23		0.2773			
Video-LDM[91]	CVPR23		10M	256×256			0.2929		550.61	33.45
Show-1[43]	Arxiv23		10M	320×576	13.08	538	0.3072		394.46	35.42
PixelDance[191]	Arxiv23		10M	336×596		381	0.3125	**49.36**	242.82	42.1
HiGen[64]	Arxiv23		30M	448×256	8.60	406	0.2947			
VersVideo[42]	ICLR24		10M	256×256		421	0.3014		**119**	**81.3**
MicroCinema[192]	Arxiv23		10M	448×448		377.4	0.2967		342.86	37.46
VideoComposer[69]	NeurIPS23		10M	256×256		580	0.2932			
UniVG[193]	Arxiv24		11M	1280×720		336	0.3014			
TF-T2V[60]	Arxiv23		20M	448×256	**8.19**	441	0.2991			
ART•V[54]	Arxiv23		5M	768×768		**291.08**	0.2859		315.69	50.34
MoVideo[45]	Arxiv23		10M	256×256	12.71		**0.3213**		313.41	34.13
Make-A-Video[5]	Arxiv22	WebVid[122] HD-VILA[147]	20M	256×256	13.17		0.3049		367.23	33
MagicVideo[92]	Arxiv22		20M	256×256	36.50	998		145	655	
VideoFactory[15]	Arxiv24			256×256			0.3005		410	
InternVid[10]	Arxiv23	WebVid[122] InternVid[10]	28M	256×256			0.2951	60.25	616.51	21.04
VidRD[51]	Arxiv23	WebVid[122] Kinetics[158] VideoLT[194]	5.3M	256×256					363.19	39.37
LaVie[44]	Arxiv23	WebVid[122] Vimeo[44]	35M	320×512			0.2949		526.3	
Imagen Video[89]	Arxiv23		14M	1280×768						
FusionFrames[195]	Arxiv23			256×256			0.2976		433.05	24.33
W.A.L.T[58]	Arxiv23		89M	128 × 224					258.1	35.1

VIII. Conclusion

In this article, we present a thorough survey of text-to-video generation techniques and systematically categorize methods into 1) VAE-based approach, 2) GAN-based approach, 3) Auto-regressive transformer based approach, 4) Diffusion-based approach, 5) Diffusion model-based approach, and 6) T2I for video generation approach. This survey comprehensively reviews nearly ninety representative T2V generation approaches and includes the latest method published in March 2024. In addition, we introduce 40 video datasets, 20 evaluation metrics, and available open source T2V models, making it easy for the reader who would like to work on T2V generation research. Furthermore, we report comparative performance evaluations. In the end, we discuss challenges and future trends that move the field forward.

Acknowledgements

This work was supported by the National Natural Science Foundation of China No.62206123, xxx

References

[1] A. Ramesh, M. Pavlov, G. Goh, S. Gray, *et al.*, "Zero-shot text-to-image generation", in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, *Proceedings of Machine Learning Research*, vol.139, pp.8821–8831, https://proceedings.mlr.press/v139/ramesh21a.html, 2021.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents", *arXiv preprint arXiv:2204.06125*, vol.1, no.2, artilce no.3, 2022.

[3] M. Ding, W. Zheng, W. Hong, and J. Tang, "Cogview2: Faster and better text-to-image generation via hierarchical transformers", in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol.35, pp.16890–16902, https://proceedings.neurips.cc/paper_files/paper/2022/file/6baec7c4ba0a8734ccbd528a8090cb1f-Paper-Conference.pdf, 2022.

[4] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers", 2022, 2205.15868.

[5] U. Singer, A. Polyak, T. Hayes, X. Yin, *et al.*, "Make-a-video: Text-to-video generation without text-video data", 2022, 2209.14792.

[6] J. Ho, T. Salimans, A. Gritsenko, W. Chan, *et al.*, "Video diffusion models", in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol.35, pp.8633–8646, https://proceedings.neurips.cc/paper_files/paper/2022/file/39235c56aef13fb05a6adc95eb9d8d66-Paper-Conference.pdf, 2022.

[7] M. Chen, X. Tan, B. Li, Y. Liu, *et al.*, "Adaspeech: Adaptive text to speech for custom voice", 2021, 2103.00993.

[8] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, *et al.*, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning", 2019, 1907.04448.

[9] D. Paul, M. P. Shifas, Y. Pantazis, and Y. Stylianou, "Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion", 2020, 2008.05809.

[10] Y. Wang, Y. He, Y. Li, K. Li, *et al.*, "Internvid: A large-scale video-text dataset for multimodal understanding and generation", *arXiv preprint arXiv:2307.06942*, in press, 2023.

[11] T. Brooks, B. Peebles, C. Holmes, W. DePue, *et al.*, "Video generation models as world simulators", in press, https://openai.com/research/video-generation-models-as-world-simulators, 2024.

[12] A. Singh, "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, doi:10.1109/airc57904.2023.10303174, http://dx.doi.org/10.1109/AIRC57904.2023.10303174, 2023.

[13] J. Cho, F. D. Puspitasari, S. Zheng, J. Zheng, *et al.*, "Sora as an agi world model? a complete survey on text-to-video generation", 2024, 2403.05131.

[14] Z. Xing, Q. Feng, H. Chen, Q. Dai, *et al.*, "A survey on video diffusion models", 2023, 2310.10647.

[15] W. Wang, H. Yang, Z. Tuo, H. He, *et al.*, "Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation", *arXiv preprint arXiv:2305.10874*, in press, 2023.

[16] Y. Liu, X. Cun, X. Liu, X. Wang, *et al.*, "Evalcrafter: Benchmarking and evaluating large video generation models", in press, 2023.

[17] Y. Liu, L. Li, S. Ren, R. Gao, *et al.*, "Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation", in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol.36, pp.62352–62387, https://proceedings.neurips.cc/paper_files/paper/2023/file/c481049f7410f38e788f67c171c64ad5-Paper-Datasets_and_Benchmarks.pdf, 2023.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, *et al.*, "Learning transferable visual models from natural language supervision", in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, *Proceedings of Machine Learning Research*, vol.139, pp.8748–8763, https://proceedings.mlr.press/v139/radford21a.html, 2021.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", 2019, 1810.04805.

[20] C. Raffel, N. Shazeer, A. Roberts, K. Lee, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer", *Journal of Machine Learning Research*, vol.21, no.140, pp.1–67, http://jmlr.org/papers/v21/20-074.html, 2020.

[21] H. Touvron, T. Lavril, G. Izacard, X. Martinet, *et al.*, "Llama: Open and efficient foundation language models", 2023, 2302.13971.

[22] D. P. Kingma and M. Welling, "Auto-encoding variational bayes", 2022, 1312.6114.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, *et al.*, "Generative adversarial networks", *Commun. ACM*, vol.63, no.11, artilce no.139–144, ISSN 0001-0782, doi:10.1145/3422622, https://doi.org/10.1145/3422622, 2020.

[24] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models", in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol.33, pp.6840–6851, https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf, 2020.

[25] G. Mittal, T. Marwah, and V. N. Balasubramanian, "Sync-draw: Automatic video generation using deep recurrent attentive architectures", in *Proceedings of the 25th ACM International Conference on Multimedia*, Association for Computing Machinery, New York, NY, USA, MM '17, artilce no.1096–1104, doi:10.1145/3123266.3123309, https://doi.org/10.1145/3123266.3123309, 2017, ISBN 9781450349062.

[26] C. Wu, L. Huang, Q. Zhang, B. Li, *et al.*, "Godiva: Generating open-domain videos from natural descriptions", 2021, 2104.14806.

[27] Y. Pan, Z. Qiu, T. Yao, H. Li, and T. Mei, "To create what you tell: Generating videos from captions", in *Proceedings of the 25th ACM International Conference on Multimedia*, Association for Computing Machinery, New York, NY, USA, MM '17, artilce no.1789–1798, doi:10.1145/3123266.3127905, https://doi.org/10.1145/3123266.3127905, 2017, ISBN 9781450349062.

[28] Y. Balaji, M. R. Min, B. Bai, R. Chellappa, and H. P. Graf, "Conditional gan with discriminative filter generation for text-to-video synthesis.", in *IJCAI*, vol.1, artilce no.2, 2019.

[29] Y. Li, M. Min, D. Shen, D. Carlson, and L. Carin, "Video generation from text", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.32, no.1, doi:10.1609/aaai.v32i1.12233, https://ojs.aaai.org/index.php/AAAI/article/view/12233, 2018.

[30] K. Deng, T. Fei, X. Huang, and Y. Peng, "Irc-gan: Introspective recurrent convolutional gan for text-to-video generation.", in *IJCAI*, pp.2216–2222, 2019.

[31] D. Kim, D. Joo, and J. Kim, "Tivgan: Text to image to video generation with step-by-step evolutionary generator", *IEEE Access*, vol.8, pp.153113–153122, doi:10.1109/ACCESS.2020.3017881, 2020.

[32] Y. Li, Z. Gan, Y. Shen, J. Liu, *et al.*, "Storygan: A sequential conditional gan for story visualization", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] B. Li, "Word-level fine-grained story visualization", in *Computer Vision – ECCV 2022*, Springer Nature Switzerland, Cham, pp.347–362, 2022, ISBN 978-3-031-20059-5.

[34] C. Wu, J. Liang, L. Ji, F. Yang, *et al.*, "Nüwa: Visual synthesis pre-training for neural visual world creation", in *Computer Vision – ECCV 2022*, Springer Nature Switzerland, Cham, pp.720–736, 2022, ISBN 978-3-031-19787-1.

[35] D. Kondratyuk, L. Yu, X. Gu, J. Lezama, *et al.*, "Videopoet: A large language model for zero-shot video generation", 2024, 2312.14125.

[36] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, *et al.*, "Phenaki: Variable length video generation from open domain textual descriptions", in *International Conference on Learning Representations*, https://openreview.net/forum?id=vOEXS39nOF, 2023.

[37] L. Han, J. Ren, H.-Y. Lee, F. Barbieri, *et al.*, "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 (CVPR)*, pp.3615–3625, 2022.

[38] H. Liu, W. Yan, M. Zaharia, and P. Abbeel, "World model on million-length video and language with blockwise ringattention", 2024, 2402.08268.

[39] L. Yu, Y. Cheng, K. Sohn, J. Lezama, *et al.*, "Magvit: Masked generative video transformer", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.10459–10469, 2023.

[40] X. Wang, Z. Zhu, G. Huang, B. Wang, *et al.*, "Worlddreamer: Towards general world models for video generation via predicting masked tokens", 2024, 2401.09985.

[41] Y. He, T. Yang, Y. Zhang, Y. Shan, and Q. Chen, "Latent video diffusion models for high-fidelity long video generation", 2023, 2211.

Chinese Journal of Electronics

13221.

[42] J. Xiang, R. Huang, J. Zhang, G. Li, *et al.*, "Versvideo: Leveraging enhanced temporal diffusion models for versatile video generation", in *The Twelfth International Conference on Learning Representations*, https://openreview.net/forum?id=K9sVJ17zvB, 2024.

[43] D. J. Zhang, J. Z. Wu, J.-W. Liu, R. Zhao, *et al.*, "Show-1: Marrying pixel and latent diffusion models for text-to-video generation", 2023, 2309.15818.

[44] Y. Wang, X. Chen, X. Ma, S. Zhou, *et al.*, "Lavie: High-quality video generation with cascaded latent diffusion models", 2023, 2309.15103.

[45] J. Liang, Y. Fan, K. Zhang, R. Timofte, *et al.*, "Movideo: Motion-aware video generation with diffusion models", 2023, 2311.11325.

[46] Z. Yuan, R. Chen, Z. Li, H. Jia, *et al.*, "Mora: Enabling generalist video generation via a multi-agent framework", 2024, 2403.13248.

[47] Y. Zhang, Y. Wei, X. Lin, Z. Hui, *et al.*, "Videoelevator: Elevating video generation quality with versatile text-to-image diffusion models", 2024, 2403.05438.

[48] Z. Luo, D. Chen, Y. Zhang, Y. Huang, *et al.*, "Videofusion: Decomposed diffusion models for high-quality video generation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.10209–10218, 2023.

[49] H. Yuan, S. Zhang, X. Wang, Y. Wei, *et al.*, "Instructvideo: Instructing video diffusion models with human feedback", 2023, 2312.12490.

[50] S. Ma, H. Xu, M. Li, W. Geng, *et al.*, "Pos: A prompts optimization suite for augmenting text-to-video generation", 2024, 2311.00949.

[51] J. Gu, S. Wang, H. Zhao, T. Lu, *et al.*, "Reuse and diffuse: Iterative denoising for text-to-video generation", 2023, 2309.03549.

[52] S. Su, J. Liu, L. Gao, and J. Song, "F³-pruning: A training-free and generalized pruning strategy towards faster and finer text-to-video synthesis", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.38, no.5, pp.4961–4969, doi:10.1609/aaai.v38i5.28300, https://ojs.aaai.org/index.php/AAAI/article/view/28300, 2024.

[53] X. Wang, S. Zhang, H. Zhang, Y. Liu, *et al.*, "Videolcm: Video latent consistency model", 2023, 2312.09109.

[54] W. Weng, R. Feng, Y. Wang, Q. Dai, *et al.*, "Art·v: Auto-regressive text-to-video generation with diffusion models", 2023, 2311.18834.

[55] H. Zhang, Z. Wu, Z. Xing, J. Shao, and Y.-G. Jiang, "Adadiff: Adaptive step selection for fast diffusion", 2023, 2311.14768.

[56] H. Lu, G. Yang, N. Fei, Y. Huo, *et al.*, "VDT: General-purpose video diffusion transformers via mask modeling", in *The Twelfth International Conference on Learning Representations*, https://openreview.net/forum?id=Un0rgm9f04, 2024.

[57] X. Ma, Y. Wang, G. Jia, X. Chen, *et al.*, "Latte: Latent diffusion transformer for video generation", *arXiv preprint arXiv:2401.03048*, in press, 2024.

[58] A. Gupta, L. Yu, K. Sohn, X. Gu, *et al.*, "Photorealistic video generation with diffusion models", 2023, 2312.06662.

[59] W. Menapace, A. Siarohin, I. Skorokhodov, E. Deyneka, *et al.*, "Snap video: Scaled spatiotemporal transformers for text-to-video synthesis", 2024, 2402.14797.

[60] X. Wang, S. Zhang, H. Yuan, Z. Qing, *et al.*, "A recipe for scaling up text-to-video generation with text-free videos", in *CVPR*, 2024.

[61] L. Hu, X. Gao, P. Zhang, K. Sun, *et al.*, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation", *arXiv preprint arXiv:2311.17117*, in press, 2023.

[62] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.7346–7356, 2023.

[63] J. Xing, M. Xia, Y. Zhang, H. Chen, *et al.*, "Dynamicrafter: Animating open-domain images with video diffusion priors", 2023, 2310.12190.

[64] Z. Qing, S. Zhang, J. Wang, X. Wang, *et al.*, "Hierarchical spatio-temporal decoupling for text-to-video generation", 2023, 2312.04483.

[65] R. Wu, L. Chen, T. Yang, C. Guo, *et al.*, "Lamp: Learn a motion pattern for few-shot-based video generation", 2023, 2310.10769.

[66] R. Zhao, Y. Gu, J. Z. Wu, D. J. Zhang, *et al.*, "Motiondirector: Motion customization of text-to-video diffusion models", 2023, 2310.08465.

[67] Y. Jain, A. Nasery, V. Vineet, and H. Behl, "Peekaboo: Interactive video generation via masked-diffusion", *arXiv preprint arXiv:2312.07509*, in press, 2023.

[68] Y. Zhang, Y. Wei, D. Jiang, X. Zhang, *et al.*, "Controlvideo: Training-free controllable text-to-video generation", 2023, 2305.13077.

[69] X. Wang, H. Yuan, S. Zhang, D. Chen, *et al.*, "Videocomposer: Compositional video synthesis with motion controllability", in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol.36, pp.7594–7611, https://proceedings.neurips.cc/paper_files/paper/2023/file/180f6184a3458fa19c28c5483bc61877-Paper-Conference.pdf, 2023.

[70] G. Liu, M. Xia, Y. Zhang, H. Chen, *et al.*, "Stylecrafter: Enhancing stylized text-to-video generation with style adapter", 2023, 2312.00330.

[71] J. Zhu, H. Yang, W. Wang, H. He, *et al.*, "Mobilevidfactory: Automatic diffusion-based social media video generation for mobile devices from text", 2023, 2307.16371.

[72] J. Wang, Y. Zhang, J. Zou, Y. Zeng, *et al.*, "Boximator: Generating rich and controllable motions for video synthesis", 2024, 2402.01566.

[73] Z. Duan, L. You, C. Wang, C. Chen, *et al.*, "Diffsynth: Latent iteration deflickering for realistic video synthesis", 2023, 2308.03463.

[74] B. Liu, X. Liu, A. Dai, Z. Zeng, *et al.*, "Dual-stream diffusion net for text-to-video generation", 2023, 2308.08316.

[75] H. Fei, S. Wu, W. Ji, H. Zhang, and T.-S. Chua, "Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms", 2024, 2308.13812.

[76] H. Lin, A. Zala, J. Cho, and M. Bansal, "Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning", 2023, 2309.15091.

[77] L. Lian, B. Shi, A. Yala, T. Darrell, and B. Li, "Llm-grounded video diffusion models", 2023, 2309.17444.

[78] Y. Jiang, S. Yang, T. L. Koh, W. Wu, *et al.*, "Text2performer: Text-driven human video generation", 2023, 2304.08483.

[79] M. Yang, Y. Du, B. Dai, D. Schuurmans, *et al.*, "Probabilistic adaptation of text-to-video models", 2023, 2306.01872.

[80] X. Li, Y. Zhang, and X. Ye, "Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model", 2023, 2310.07771.

[81] S. Yin, C. Wu, H. Yang, J. Wang, *et al.*, "Nuwa-xl: Diffusion over diffusion for extremely long video generation", 2023, 2303.12346.

[82] X. Chen, Y. Wang, L. Zhang, S. Zhuang, *et al.*, "Seine: Short-to-long video diffusion model for generative transition and prediction", 2023, 2310.20700.

[83] G. Oh, J. Jeong, S. Kim, W. Byeon, *et al.*, "Mtvg : Multi-text video generation with text-to-video models", 2023, 2312.04086.

[84] H. Qiu, M. Xia, Y. Zhang, Y. He, *et al.*, "Freenoise: Tuning-free longer video diffusion via noise rescheduling", 2024, 2310.15169.

[85] F.-Y. Wang, W. Chen, G. Song, H.-J. Ye, *et al.*, "Gen-l-video: Multi-text to long video generation via temporal co-denoising", 2023, 2305.18264.

[86] R. Henschel, L. Khachatryan, D. Hayrapetyan, H. Poghosyan, *et al.*, "Streamingt2v: Consistent, dynamic, and extendable long video generation from text", 2024, 2403.14773.

[87] S. Zhuang, K. Li, X. Chen, Y. Wang, *et al.*, "Vlogger: Make your dream a vlog", 2024, 2401.09414.

[88] J. Wang, H. Yuan, D. Chen, Y. Zhang, *et al.*, "Modelscope text-to-video technical report", 2023, 2308.06571.

Chinese Journal of Electronics

[89] J. Ho, W. Chan, C. Saharia, J. Whang, *et al.*, "Imagen video: High definition video generation with diffusion models", 2022, 2210.02303.

[90] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, *et al.*, "Lumiere: A space-time diffusion model for video generation", 2024, 2401.12945.

[91] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, *et al.*, "Align your latents: High-resolution video synthesis with latent diffusion models", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.22563–22575, 2023.

[92] D. Zhou, W. Wang, H. Yan, W. Lv, *et al.*, "Magicvideo: Efficient video generation with latent diffusion models", 2023, 2211.11018.

[93] Z. Xing, Q. Dai, H. Hu, Z. Wu, and Y.-G. Jiang, "Simda: Simple diffusion adapter for efficient video generation", 2023, 2308.09710.

[94] W. Wang, J. Liu, Z. Lin, J. Yan, *et al.*, "Magicvideo-v2: Multi-stage high-aesthetic video generation", 2024, 2401.04468.

[95] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets", 2023, 2311.15127.

[96] X. Li, W. Chu, Y. Wu, W. Yuan, *et al.*, "Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation", 2023, 2309.00398.

[97] S. Ge, S. Nah, G. Liu, T. Poon, *et al.*, "Preserve your own correlation: A noise prior for video diffusion models", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.22930–22941, 2023.

[98] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, *et al.*, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.15954–15964, 2023.

[99] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, *et al.*, "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 (ICCV)*, pp.7623–7633, 2023.

[100] J. An, S. Zhang, H. Yang, S. Gupta, *et al.*, "Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation", 2023, 2304.08477.

[101] H. Chen, Y. Zhang, X. Cun, M. Xia, *et al.*, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models", 2024, 2401.09047.

[102] T. Lee, S. Kwon, and T. Kim, "Grid diffusion models for text-to-video generation", 2024, 2404.00234.

[103] S. Hong, J. Seo, H. Shin, S. Hong, and S. Kim, "Direct2v: Large language models are frame-level directors for zero-shot text-to-video generation", 2024, 2305.14330.

[104] H. Huang, Y. Feng, C. Shi, L. Xu, *et al.*, "Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator", 2023, 2309.14494.

[105] Y. Lu, L. Zhu, H. Fan, and Y. Yang, "Flowzero: Zero-shot text-to-video synthesis with llm-driven dynamic scene syntax", 2023, 2311.15813.

[106] J. Lv, Y. Huang, M. Yan, J. Huang, *et al.*, "Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning", 2024, 2311.12631.

[107] A. van den Oord, O. Vinyals, and k. kavukcuoglu, "Neural discrete representation learning", in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol.30, https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf, 2017.

[108] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild", *Computer Speech & Language*, vol.60, artilce no.101027, ISSN 0885-2308, doi:https://doi.org/10.1016/j.csl.2019.101027, https://www.sciencedirect.com/science/article/pii/S0885230819302712, 2020.

[109] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.10684–10695, 2022.

[110] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, Cham, pp.234–241, 2015, ISBN 978-3-319-24574-4.

[111] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models", 2023, 2303.01469.

[112] T. Chen and L. Li, "Fit: Far-reaching interleaved transformers", 2023, 2305.12689.

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

[114] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, *et al.*, "Gpt-4 technical report", *arXiv preprint arXiv:2303.08774*, in press, 2023.

[115] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.3836–3847, 2023.

[116] C. Saharia, W. Chan, S. Saxena, L. Li, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding", 2022, 2205.11487.

[117] Y. Balaji, S. Nah, X. Huang, A. Vahdat, *et al.*, "Ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers", 2023, 2211.01324.

[118] Y. Ma, Y. He, X. Cun, X. Wang, *et al.*, "Follow your pose: Pose-guided text-to-video generation using pose-free videos", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.38, no.5, pp.4117–4125, doi:10.1609/aaai.v38i5.28206, https://ojs.aaai.org/index.php/AAAI/article/view/28206, 2024.

[119] B. Peng, X. Chen, Y. Wang, C. Lu, and Y. Qiao, "Conditionvideo: Training-free condition-guided video generation", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.38, no.5, pp.4459–4467, doi:10.1609/aaai.v38i5.28244, https://ojs.aaai.org/index.php/AAAI/article/view/28244, 2024.

[120] F. Shi, J. Gu, H. Xu, S. Xu, *et al.*, "Bivdiff: A training-free framework for general-purpose video synthesis via bridging image and video diffusion models", 2024, 2312.02813.

[121] Y. Guo, C. Yang, A. Rao, Z. Liang, *et al.*, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning", 2024, 2307.04725.

[122] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.1728–1738, 2021.

[123] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, *et al.*, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[124] J. C. Stroud, Z. Lu, C. Sun, J. Deng, *et al.*, "Learning video representations from textual web supervision", 2021, 2007.14937.

[125] A. Nagrani, P. H. Seo, B. Seybold, A. Hauth, *et al.*, "Learning audio-video modalities from image captions", in *Computer Vision – ECCV 2022*, Springer Nature Switzerland, Cham, pp.407–426, 2022, ISBN 978-3-031-19781-9.

[126] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, pp.2556–2565, doi:10.18653/v1/P18-1238, https://aclanthology.org/P18-1238, 2018.

[127] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation", in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, *Proceedings of Machine Learning Research*, vol.162, pp.12888–12900, https://proceedings.

Chinese Journal of Electronics

mlr.press/v162/li22n.html, 2022.

[128] X. Huang, Y. Zhang, J. Ma, W. Tian, *et al.*, "Tag2text: Guiding vision-language model via image tagging", 2024, 2303.05657.

[129] K. Li, Y. He, Y. Wang, Y. Li, *et al.*, "Videochat: Chat-centric video understanding", 2024, 2305.06355.

[130] T.-S. Chen, A. Siarohin, W. Menapace, E. Deyneka, *et al.*, "Panda-70m: Captioning 70m videos with multiple cross-modality teachers", 2024, 2402.19479.

[131] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding", 2023, 2306.02858.

[132] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models", 2023, 2304.10592.

[133] K. Li, Y. Wang, Y. Li, Y. Wang, *et al.*, "Unmasked teacher: Towards training-efficient video foundation models", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.19948–19960, 2023.

[134] W. Wang, Y. Sun, and Y. Yang, "Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models", 2024, 2403.06098.

[135] Z. Yang, L. Li, K. Lin, J. Wang, *et al.*, "The dawn of lmms: Preliminary explorations with gpt-4v (ision)", *arXiv preprint arXiv:2309.17421*, vol.9, no.1, artilce no.1, 2023.

[136] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation", in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, Portland, OR, 2011.

[137] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[138] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, *et al.*, "Localizing moments in video with natural language", in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[139] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, *et al.*, "Movie description", *International Journal of Computer Vision*, vol.123, pp.94–120, 2017.

[140] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos", in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[141] L. Zhou, C. Xu, and J. Corso, "Towards automatic learning of procedures from web instructional videos", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.32, no.1, doi:10.1609/aaai.v32i1.12342, https://ojs.aaai.org/index.php/AAAI/article/view/12342, 2018.

[142] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, *et al.*, "How2: A large-scale dataset for multimodal language understanding", 2018, 1811.00347.

[143] X. Wang, J. Wu, J. Chen, L. Li, *et al.*, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[144] R. Zellers, X. Lu, J. Hessel, Y. Yu, *et al.*, "Merlot: Multimodal neural script knowledge models", in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol.34, pp.23634–23651, https://proceedings.neurips.cc/paper_files/paper/2021/file/c6d4eb15f1e84a36eff58eca3627c82e-Paper.pdf, 2021.

[145] H. Reynaud, M. Qiao, M. Dombrowski, T. Day, *et al.*, "Feature-conditioned cascaded video diffusion models for precise echocardiogram synthesis", in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, Springer Nature Switzerland, Cham, pp.142–152, 2023, ISBN 978-3-031-43999-5.

[146] T. Wang, L. Li, K. Lin, Y. Zhai, *et al.*, "DisCo: Disentangled Control for Realistic Human Dance Generation", *arXiv e-prints*, in press, article no.arXiv:2307.00040, doi:10.48550/arXiv.2307.00040, 2023.

[147] H. Xue, T. Hang, Y. Zeng, Y. Sun, *et al.*, "Advancing high-resolution video-language representation with large-scale video transcriptions", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5036–5045, 2022.

[148] J. Yu, H. Zhu, L. Jiang, C. C. Loy, *et al.*, "Celebv-text: A large-scale facial text-video dataset", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.14805–14814, 2023.

[149] X. Ju, "Mira: A mini-step towards sora-like long video generation", https://github.com/mira-space.

[150] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild", 2012, 1212.0402.

[151] W. Kay, J. Carreira, K. Simonyan, B. Zhang, *et al.*, "The kinetics human action video dataset", 2017, 1705.06950.

[152] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, *et al.*, "The "something something" video database for learning and evaluating visual common sense", in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[153] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, *et al.*, "The 2017 davis challenge on video object segmentation", 2018, 1704.00675.

[154] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach", in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol.3, pp.32–36 Vol.3, doi:10.1109/ICPR.2004.1334462, 2004.

[155] N. Aifanti, C. Papachristou, and A. Delopoulos, "The mug facial expression database", in *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, pp.1–4, 2010.

[156] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, *et al.*, "The cityscapes dataset for semantic urban scene understanding", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[157] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms", in *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, Lille, France, *Proceedings of Machine Learning Research*, vol.37, pp.843–852, https://proceedings.mlr.press/v37/srivastava15.html, 2015.

[158] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[159] F. Ebert, C. Finn, A. X. Lee, and S. Levine, "Self-supervised visual planning with temporal skip connections.", *CoRL*, vol.12, artilce no.16, 2017.

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

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

[162] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, *et al.*, "Moments in time dataset: One million videos for event understanding", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.42, no.2, pp.502–508, doi:10.1109/TPAMI.2019.2901464, 2020.

[163] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation", in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol.32, https://proceedings.neurips.cc/paper_files/paper/2019/file/31c0b36aef265d9221af80872ceb62f9-Paper.pdf, 2019.

[164] W. Liu, Z. Piao, Z. Tu, W. Luo, *et al.*, "Liquid warping gan with attention: A unified framework for human image synthesis", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.44, no.9, pp.5114–5132, doi:10.1109/TPAMI.2021.3078270, 2022.

[165] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, *et al.*, "Bridge data: Boosting generalization of robotic skills with cross-domain datasets", 2021, 2109.13396.

[166] T. Brooks, J. Hellsten, M. Aittala, T.-C. Wang, *et al.*, "Generating long videos of dynamic scenes", in *Advances in Neural Information Processing Systems*, Curran Associates,

Inc., vol.35, pp.31769–31781, https://proceedings.neurips.cc/paper_files/paper/2022/file/ce208d95d020b023cba9e64031db2584-Paper-Conference.pdf, 2022.

[167] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, *et al.*, "Towards accurate generative models of video: A new metric & challenges", *arXiv preprint arXiv:1812.01717*, in press, 2018.

[168] M. Saito, S. Saito, M. Koyama, and S. Kobayashi, "Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan", 2020, doi:10.1007/s11263-020-01333-y, https://doi.org/10.1007/s11263-020-01333-y.

[169] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium", in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol.30, https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf, 2017.

[170] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, *et al.*, "Going deeper with convolutions", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[171] J. Deng, W. Dong, R. Socher, L.-J. Li, *et al.*, "Imagenet: A large-scale hierarchical image database", in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp.248–255, doi:10.1109/CVPR.2009.5206848, 2009.

[172] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks", in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[173] H. Wu, E. Zhang, L. Liao, C. Chen, *et al.*, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives", in *International Conference on Computer Vision (ICCV)*, 2023.

[174] D. Podell, Z. English, K. Lacey, A. Blattmann, *et al.*, "Sdxl: Improving latent diffusion models for high-resolution image synthesis", 2023, 2307.01952.

[175] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models", in *Proceedings of the 40th International Conference on Machine Learning*, PMLR, *Proceedings of Machine Learning Research*, vol.202, pp.19730–19742, https://proceedings.mlr.press/v202/li23q.html, 2023.

[176] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation", in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp.311–318, 2002.

[177] D. Klakow and J. Peters, "Testing the correlation of word error rate and perplexity", *Speech Communication*, vol.38, no.1, pp.19–28, ISSN 0167-6393, doi:https://doi.org/10.1016/S0167-6393(01)00041-3, https://www.sciencedirect.com/science/article/pii/S0167639301000413, 2002.

[178] Y. Sun, Z. Ni, C.-K. Chng, Y. Liu, *et al.*, "Icdar 2019 competition on large-scale street view text with partial labeling - rrc-lsvt", in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp.1557–1562, doi:10.1109/ICDAR.2019.00250, 2019.

[179] A. C. Morris, V. Maier, and P. Green, "From wer and ril to mer and wil: Improved evaluation measures for connected speech recognition", in *Eighth International Conference on Spoken Language Processing*, 2004.

[180] S. I. Serengil and A. Ozpinar, "Hyperextended lightface: A facial attribute analysis framework", in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pp.1–4, doi:10.1109/ICEET53442.2021.9659697, 2021.

[181] M. Otani, R. Togashi, Y. Sawai, R. Ishigami, *et al.*, "Toward verifiable and reproducible human evaluation for text-to-image generation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.14277–14286, 2023.

[182] G. Parmar, R. Zhang, and J.-Y. Zhu, "On aliased resizing and surprising subtleties in gan evaluation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.11410–11420, 2022.

[183] C. Saharia, W. Chan, S. Saxena, L. Li, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding", in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol.35, pp.36479–36494, https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf, 2022.

[184] Z. Huang, Y. He, J. Yu, F. Zhang, *et al.*, "VBench: Comprehensive benchmark suite for video generative models", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[185] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, *et al.*, "Microsoft coco: Common objects in context", in *Computer Vision – ECCV 2014*, Springer International Publishing, Cham, pp.740–755, 2014, ISBN 978-3-319-10602-1.

[186] J. Cho, A. Zala, and M. Bansal, "Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.3043–3054, 2023.

[187] J. H. Zar, "Spearman rank correlation", in *Encyclopedia of Biostatistics*, John Wiley & Sons, Ltd, 2005, ISBN 9780470011812, doi:https://doi.org/10.1002/0470011815.b2a15150, https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470011815.b2a15150, https://onlinelibrary.wiley.com/doi/abs/10.1002/0470011815.b2a15150.

[188] M. G. Kendall, "Rank correlation methods.", in press, 1948.

[189] M. Contributors, "Openmmlab's next generation video understanding toolbox and benchmark", https://github.com/open-mmlab/mmaction2, 2020.

[190] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow", in *Computer Vision – ECCV 2020*, Springer International Publishing, Cham, pp.402–419, 2020, ISBN 978-3-030-58536-5.

[191] Y. Zeng, G. Wei, J. Zheng, J. Zou, *et al.*, "Make pixels dance: High-dynamic video generation", 2023, 2311.10982.

[192] Y. Wang, J. Bao, W. Weng, R. Feng, *et al.*, "Microcinema: A divide-and-conquer approach for text-to-video generation", 2023, 2311.18829.

[193] L. Ruan, L. Tian, C. Huang, X. Zhang, and X. Xiao, "Univg: Towards unified-modal video generation", 2024, 2401.09084.

[194] X. Zhang, Z. Wu, Z. Weng, H. Fu, *et al.*, "Videolt: Large-scale long-tailed video recognition", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.7960–7969, 2021.

[195] V. Arkhipkin, Z. Shaheen, V. Vasilev, E. Dakhova, *et al.*, "Fusionframes: Efficient architectural aspects for text-to-video generation pipeline", 2023, 2311.13073.

Firstname1 Middlename1 Lastname1 With a frequency range of approximately 70–300 MHz, the MWA spans a number of Earth and space-based broadcast bands, including the ubiquitous FM band (approximately 88–108 MHz in Australia), constituting a primary source of RFI at the MRO. Likewise, the frequency range for SKA_low is 50–350 MHz, also encompassing the FM band. With a frequency range of approximately 70–300 MHz, the MWA spans a number of Earth and space-based broadcast bands. (Email: xxxxxxxx@xxx.xxx.xx)

Firstname2 Middlename2 Lastname2 With a frequency range of approximately 70–300 MHz, the MWA spans a number of Earth and space-based broadcast bands, including the ubiquitous FM band (approximately 88–108 MHz in Australia), constituting a primary source of RFI at the MRO. Likewise, the frequency range for SKA_low is 50–350 MHz, also encompassing the FM band. With a frequency range of approximately 70–300 MHz, the MWA spans a number of Earth and space-based broadcast bands. With a frequency range of approximately 70–300 MHz, the MWA spans a number of Earth and space-based broadcast bands, including the ubiquitous FM band (approximately 88–108 MHz in Australia), constituting a primary source of RFI at the MRO. Likewise, the frequency range for SKA_low is 50–350 MHz, also encompassing the FM band. With a frequency range of approximately 70–300 MHz, the MWA spans a number of Earth and space-based broadcast bands. (Email: xxxxxxxx@xxx.xxx.xx)

Firstname3 Middlename3 Lastname3 Xxxx xxx xxx xxx xxxxxxxx xxxxxxxxx xxxxxxxx xxxxx xxxxxxxx xxxx xxxxxxxx xxxxx xxxxx xxxxxxxx xxxxx xxxxx xxxxxxxx xxxxx xxxxx xxxxxxxx xxxxx xxxxx xxxxxxxx xxxxx xxxxx xxxxxxxx xxxxx xxxxxxxx xxxxxxxx xxxxx xxxxx xxxxxxxx xxxxx xxxxx xxxxxxxx xxxxx xxxxxx xxxxxxxx xxx (Email: xxxxxxxx@xxx.xxx.xx)

Firstname4 Middlename4 Lastname4 Xxxx xxx xxx xxx xxxxxxxx xxxxxxxxx xxxxxxxx xxxxx xxxxxxxx xxxxxxxx xxxxxxxx xxx (Email: xxxxxxxx@xxx.xxx.xx)

Chinese Journal of Electronics