Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models

Yixin Liu[1*] **Kai Zhang**[1*] **Yuan Li**[1*] **Zhiling Yan**[1*] **Chujie Gao**[1*]
Ruoxi Chen[1*] **Zhengqing Yuan**[1*] **Yue Huang**[1*] **Hanchi Sun**[1*]
Jianfeng Gao[2] **Lifang He**[1] **Lichao Sun**[1†]

[1]**Lehigh University** [2]**Microsoft Research**

Abstract

Warning: This is not an official technical report from OpenAI.

Sora is a text-to-video generative AI model, released by OpenAI in February 2024. The model is trained to generate videos of realistic or imaginative scenes from text instructions and show potential in simulating the physical world. Based on public technical reports and reverse engineering, this paper presents a comprehensive review of the model's background, related technologies, applications, remaining challenges, and future directions of text-to-video AI models. We first trace Sora's development and investigate the underlying technologies used to build this "world simulator". Then, we describe in detail the applications and potential impact of Sora in multiple industries ranging from film-making and education to marketing. We discuss the main challenges and limitations that need to be addressed to widely deploy Sora, such as ensuring safe and unbiased video generation. Lastly, we discuss the future development of Sora and video generation models in general, and how advancements in the field could enable new ways of human-AI interaction, boosting productivity and creativity of video generation.

Figure 1: Sora: A Breakthrough in AI-Powered Vision Generation.

[*]Equal contributions. The order was determined by rolling dice. Chujie, Ruoxi, Yuan, Yue, and Zhengqing are visiting students in the LAIR lab at Lehigh University. The GitHub link is https://github.com/lichao-sun/SoraReview

[†]Lichao Sun is co-corresponding author: lis221@lehigh.edu

Contents

1 Introduction

Since the release of ChatGPT in November 2022, the advent of AI technologies has marked a significant transformation, reshaping interactions and integrating deeply into various facets of daily life and industry [1, 2]. Building on this momentum, OpenAI released, in February 2024, Sora, a text-to-video generative AI model that can generate videos of realistic or imaginative scenes from text prompts. Compared to previous video generation models, Sora is distinguished by its ability to produce up to 1-minute long videos with high quality while maintaining adherence to user's text instructions [3]. This progression of Sora is the embodiment of the long-standing AI research mission of equipping AI systems (or AI Agents) with the capability of understanding and interacting with the physical world in motion. This involves developing AI models that are capable of not only interpreting complex user instructions but also applying this understanding to solve real-world problems through dynamic and contextually rich simulations.

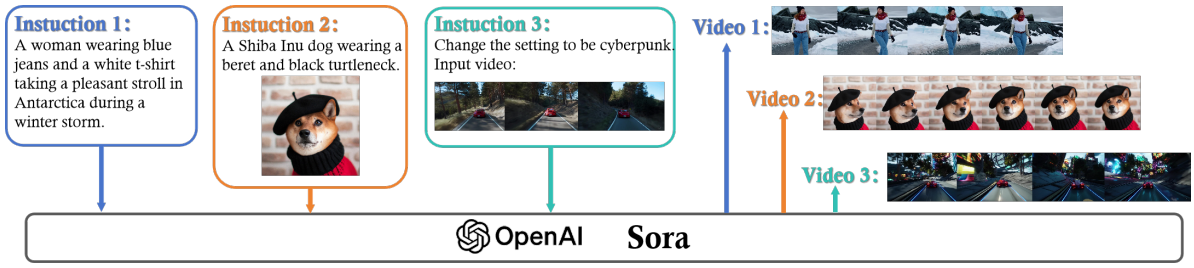


Figure 2: Examples of Sora in text-to-video generation. Text instructions are given to the OpenAI Sora model, and it generates three videos according to the instructions.

Sora demonstrates a remarkable ability to accurately interpret and execute complex human instructions, as illustrated in Figure 2. The model can generate detailed scenes that include multiple characters that perform specific actions against intricate backgrounds. Researchers attribute Sora's proficiency to not only processing user-generated textual prompts but also discerning the complicated interplay of elements within a scenario. One of the most striking aspects of Sora is its capacity for up to a minute-long video while maintaining high visual quality and compelling visual coherency. Unlike earlier models that can only generate short video clips, Sora's minute-long video creation possesses a sense of progression and a visually consistent journey from its first frame to the last. In addition, Sora's advancements are evident in its ability to produce extended video sequences with nuanced depictions of motion and interaction, overcoming the constraints of shorter clips and simpler visual renderings that characterized earlier video generation models. This capability represents a leap forward in AI-driven creative tools, allowing users to convert text narratives to rich visual stories. Overall, these advances show the potential of Sora as a *world simulator* to provide nuanced insights into the physical and contextual dynamics of the depicted scenes. [3].

Technology. At the heart of Sora is a pre-trained *diffusion transformer* [4]. Transformer models have proven scalable and effective for many natural language tasks. Similar to powerful large language models (LLMs) such as GPT-4, Sora can parse text and comprehend complex user instructions. To make video generation computationally efficient, Sora employs *spacetime latent patches* as its building blocks. Specifically, Sora compresses a raw input video into a latent spacetime representation. Then, a sequence of latent spacetime patches is extracted from the compressed video to encapsulate both the visual appearance and motion dynamics over brief intervals. These patches, analogous to word tokens in language models, provide Sora with detailed *visual phrases* to be used to construct videos. Sora's text-to-video generation is performed by a diffusion transformer model. Starting with a frame filled with visual noise, the model iteratively denoises the image and introduces specific details according to the provided text prompt. In essence, the

generated video emerges through a multi-step refinement process, with each step refining the video to be more aligned with the desired content and quality.

Highlights of Sora. Sora's capabilities have profound implications in various aspects:

- *Improving simulation abilities*: Training Sora at scale is attributed to its remarkable ability to simulate various aspects of the physical world. Despite lacking explicit 3D modeling, Sora exhibits 3D consistency with dynamic camera motion and long-range coherence that includes object persistence and simulates simple interactions with the world. Moreover, Sora intriguingly simulates digital environments like Minecraft, controlled by a basic policy while maintaining visual fidelity. These emergent abilities suggest that scaling video models is effective in creating AI models to simulate the complexity of physical and digital worlds.
- *Boosting creativity*: Imagine outlining a concept through text, whether a simple object or a full scene, and seeing a realistic or highly stylized video rendered within seconds. Sora allows an accelerated design process for faster exploration and refinement of ideas, thus significantly boosting the creativity of artists, filmmakers, and designers.
- *Driving educational innovations*: Visual aids have long been integral to understanding important concepts in education. With Sora, educators can easily turn a class plan from text to videos to captivate students' attention and improve learning efficiency. From scientific simulations to historical dramatizations, the possibilities are boundless.
- *Enhancing Accessibility*: Enhancing accessibility in the visual domain is paramount. Sora offers an innovative solution by converting textual descriptions to visual content. This capability empowers all individuals, including those with visual impairments, to actively engage in content creation and interact with others in more effective ways. Consequently, it allows for a more inclusive environment where everyone has the opportunity to express his or her ideas through videos.
- *Fostering emerging applications*: The applications of Sora are vast. For example, marketers might use it to create dynamic advertisements tailored to specific audience descriptions. Game developers might use it to generate customized visuals or even character actions from player narratives.

Limitations and Opportunities. While Sora's achievements highlight significant advancements in AI, challenges remain. Depicting complex actions or capturing subtle facial expressions are among the areas where the model could be enhanced. In addition, ethical considerations such as mitigating biases in generated content and preventing harmful visual outputs underscore the importance of responsible usage by developers, researchers, and the broader community. Ensuring that Sora's outputs are consistently safe and unbiased is a principal challenge. The field of video generation is advancing swiftly, with academic and industry research teams making relentless strides. The advent of competing text-to-video models suggests that Sora may soon be part of a dynamic ecosystem. This collaborative and competitive environment fosters innovation, leading to improved video quality and new applications that help improve the productivity of workers and make people's lives more entertaining.

Our Contributions. Based on published technical reports and our reverse engineering, this paper presents the first comprehensive review of Sora's background, related technologies, emerging applications, current limitations, and future opportunities.

2 Background

2.1 History

In the realm of computer vision (CV), prior to the deep learning revolution, traditional image generation techniques relied on methods like texture synthesis [5] and texture mapping [6], based on hand-crafted features. However, these methods were limited in their capacity to produce complex and vivid images.

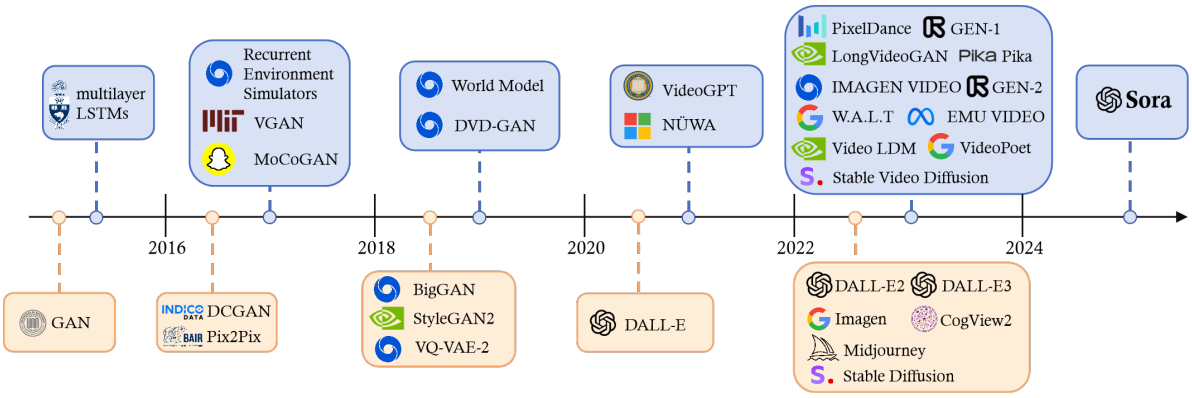


Figure 3: History of Generative AI in Vision Domain.

The introduction of Generative Adversarial Networks (GANs) [7] and Variational Autoencoders (VAEs) [8] marked a significant turning point due to its remarkable capabilities across various applications. Subsequent developments, such as flow models [9] and diffusion models [10], further enhanced image generation with greater detail and quality. The recent progress in Artificial Intelligence Generated Content (AIGC) technologies has democratized content creation, enabling users to generate desired content through simple textual instructions [11].

Over the past decade, the development of generative CV models has taken various routes, as shown in Figure 3. This landscape began to shift notably following the successful application of the transformer architecture [12] in NLP, as demonstrated by BERT [13] and GPT [14]. In CV, researchers take this concept even further by combining the transformer architecture with visual components, allowing it to be applied to downstream CV tasks, such as Vision Transformer (ViT) [15] and Swin Transformer [16]. Parallel to the transformer's success, diffusion models have also made significant strides in the fields of image and video generation [10]. Diffusion models offer a mathematically sound framework for converting noise into images with U-Nets [17], where U-Nets facilitate this process by learning to predict and mitigate noise at each step. Since 2021, a paramount focus in AI has been on generative language and vision models that are capable of interpreting human instructions, known as multimodal models. For example, CLIP [18] is a pioneering vision-language model that combines transformer architecture with visual elements, facilitating its training on vast datasets of text and images. By integrating visual and linguistic knowledge from the outset, CLIP can function as an image encoder within multimodal generation frameworks. Another notable example is Stable Diffusion [19], a versatile text-to-image AI model celebrated for its adaptability and ease of use. It employs transformer architecture and latent diffusion techniques to decode textual inputs and produce images of a wide array of styles, further illustrating the advancements in multimodal AI.

Following the release of ChatGPT in November 2022, we have witnessed the emergence of commercial text-to-image products in 2023, such as Stable Diffusion [19], Midjourney [20], DALL-E 3 [21]. These tools enable users to generate new images of high resolution and quality with simple text prompts, showcasing the potential of AI in creative image generation. However, transitioning from text-to-image to text-to-video is challenging due to the temporal complexity of videos. Despite numerous efforts in industry and academia, most existing video generation tools, such as Pika [22] and Gen-2 [23], are limited to producing only short video clips of a few seconds. In this context, `Sora` represents a significant breakthrough, akin to ChatGPT's impact in the NLP domain. `Sora` is the first model that is capable of generating videos up to one minute long based on human instructions, marking a milestone that profoundly influences research and development in generative AI. To facilitate easy access to the latest advancements in vision generation models, the most recent works have been compiled and provided in the Appendix and our GitHub.

2.2 Advanced Concepts

Scaling Laws for Vision Models. With scaling laws for LLMs, it is natural to ask whether the development of vision models follows similar scaling laws. Recently, Zhai et al. [24] have demonstrated that the performance-compute frontier for ViT models with enough training data roughly follows a (saturating) power law. Following them, Google Research [25] presented a recipe for highly efficient and stable training of a 22B-parameter ViT. Results show that great performance can be achieved using the frozen model to produce embeddings, and then training thin layers on top. `Sora`, as a large vision model (LVM), aligns with these scaling principles, uncovering several emergent abilities in text-to-video generation. This significant progression underscores the potential for LVMs to achieve advancements like those seen in LLMs.

Emergent Abilities. Emergent abilities in LLMs are sophisticated behaviors or functions that manifest at certain scales—often linked to the size of the model's parameters—that were not explicitly programmed or anticipated by their developers. These abilities are termed "emergent" because they emerge from the model's comprehensive training across varied datasets, coupled with its extensive parameter count. This combination enables the model to form connections and draw inferences that surpass mere pattern recognition or rote memorization. Typically, the emergence of these abilities cannot be straightforwardly predicted by extrapolating from the performance of smaller-scale models. While numerous LLMs, such as ChatGPT and GPT-4, exhibit emergent abilities, vision models demonstrating comparable capabilities have been scarce until the advent of Sora. According to `Sora`'s technical report, it is the first vision model to exhibit confirmed emergent abilities, marking a significant milestone in the field of computer vision.

In addition to its emergent abilities, `Sora` exhibits other notable capabilities, including instruction following, visual prompt engineering, and video understanding. These aspects of `Sora`'s functionality represent significant advancements in the vision domain and will be explored and discussed in the rest sections.

3 Technology

3.1 Overview of `Sora`

In the core essence, `Sora` is a diffusion transformer [4] with flexible sampling dimensions as shown in Figure 4. It has three parts: (1) A time-space compressor first maps the original video into latent space. (2) A ViT then processes the tokenized latent representation and outputs the denoised latent representation. (3) A CLIP-like [26] conditioning mechanism receives LLM-augmented user instructions and potentially visual prompts to guide the diffusion model to generate styled or themed videos. After many denoising

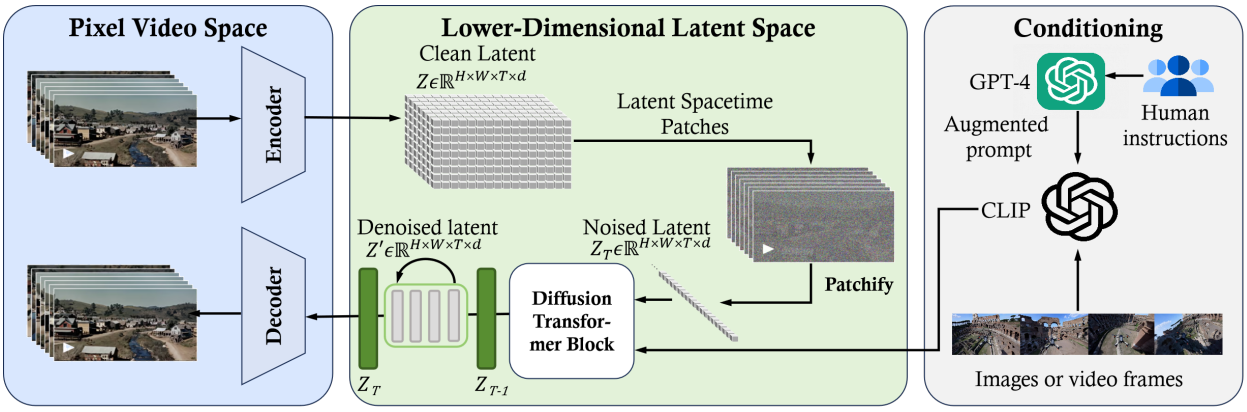


Figure 4: **Reverse Engineering**: Overview of `Sora` framework

steps, the latent representation of the generated video is obtained and then mapped back to pixel space with the corresponding decoder. In this section, we aim to reverse engineer the technology used by Sora and discuss a wide range of related works.

3.2 Data Pre-processing

3.2.1 Variable Durations, Resolutions, Aspect Ratios

One distinguishing feature of Sora is its ability to train on, understand, and generate videos and images at their native sizes [3] as illustrated in Figure 5. Traditional methods often resize, crop, or adjust the aspect ratios of videos to fit a uniform standard—typically short clips with square frames at fixed low resolutions [27][28][29]. Those samples are often generated at a wider temporal stride and rely on separately trained frame-insertion and resolution-rendering models as the final step, creating inconsistency across the video. Utilizing the diffusion transformer architecture [4] (see Section 3.2.4), Sora is the first model to embrace the diversity of visual data and can sample in a wide array of video and image formats, ranging from widescreen 1920x1080p videos to vertical 1080x1920p videos and everything in between without compromising their original dimensions.

precisely, and that is so LAME…

(a) Vertical (b) Square (c) Horizontal

Figure 5: Sora can generate images in flexible sizes or resolutions ranging from 1920x1080p to 1080x1920p and anything in between.

Training on data in their native sizes significantly improves composition and framing in the generated videos. Empirical findings suggest that by maintaining the original aspect ratios, Sora achieves a more natural and coherent visual narrative. The comparison between Sora and a model trained on uniformly cropped square videos demonstrates a clear advantage as shown in Figure 6. Videos produced by Sora exhibit better framing, ensuring subjects are fully captured in the scene, as opposed to the sometimes truncated views resulting from square cropping.

(a) Training on videos that are cropped to squares leads to unnatural compositions and framing.

(b) Training in native sizes improves framing.

Figure 6: A comparison between Sora (right) and a modified version of the model (left), which crops videos to square shapes—a common practice in model training—highlights the advantages.

This nuanced understanding and preservation of original video and image characteristics mark a significant advancement in the field of generative models. Sora's approach not only showcases the potential for more authentic and engaging video generation but also highlights the importance of diversity in training data for achieving high-quality results in generative AI. The training approach of Sora aligns with

the core tenet of Richard Sutton's THE BITTER LESSON[30], which states that leveraging computation over human-designed features leads to more effective and flexible AI systems. Just as the original design of diffusion transformers seeks simplicity and scalability [31], Sora's strategy of training on data at their native sizes eschews traditional AI reliance on human-derived abstractions, favoring instead a generalist method that scales with computational power. In the rest of this section, we try to reverse engineer the architecture design of Sora and discuss related technologies to achieve this amazing feature.

3.2.2 Unified Visual Representation

To effectively process diverse visual inputs including images and videos with varying durations, resolutions, and aspect ratios, a crucial approach involves transforming all forms of visual data into a unified representation, which facilitates the large-scale training of generative models. Specifically, Sora patchifies videos by initially compressing videos into a lower-dimensional latent space, followed by decomposing the representation into spacetime patches. However, Sora's technical report [3] merely presents a high-level idea, making reproduction challenging for the research community. In this section, we try to reverse-engineer the potential ingredients and technical pathways. Additionally, we will discuss viable alternatives that could replicate Sora's functionalities, drawing upon insights from existing literature.

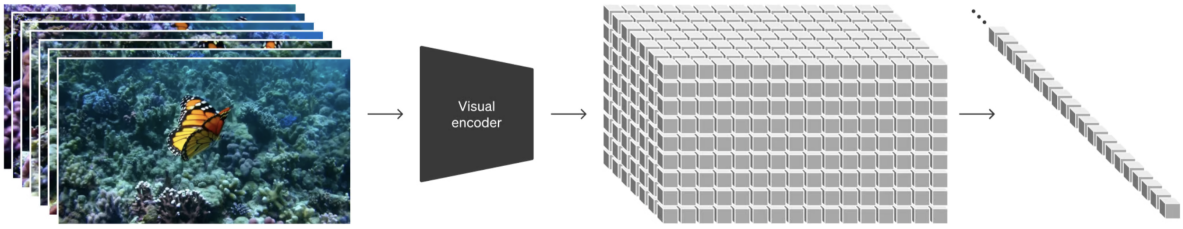


Figure 7: At a high level, Sora turns videos into patches by first compressing videos into a lower-dimensional latent space, and subsequently decomposing the representation into spacetime patches. Source: Sora's technical report [3].

3.2.3 Video Compression Network

Sora's video compression network (or visual encoder) aims to reduce the dimensionality of input data, especially a raw video, and output a latent representation that is compressed both temporally and spatially as shown in Figure 7. According to the references in the technical report, the compression network is built upon VAE or Vector Quantised-VAE (VQ-VAE) [32]. However, it is challenging for VAE to map visual data of any size to a unified and fixed-sized latent space if resizing and cropping are not used as mentioned in the technical report. We summarize two distinct implementations to address this issue:

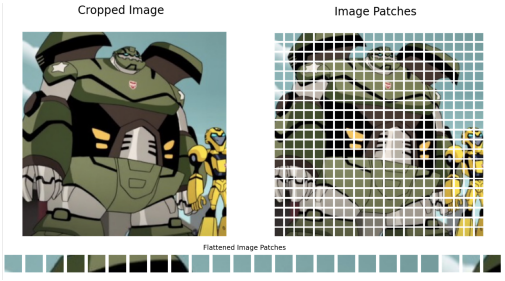


Figure 8: ViT splits an image into fixed-size patches, linearly embeds each of them, adds position embeddings, and feeds the resulting sequence of vectors to a standard Transformer encoder.

Spatial-patch Compression. This involves transforming video frames into fixed-size patches, akin to the methodologies employed in ViT [15] and MAE [33] (see Figure 8), before encoding them into a latent space. This approach is particularly effective for accommodating videos of varying resolutions and aspect ratios, as it encodes entire frames through the processing of individual patches. Subsequently, these spatial tokens are organized in a temporal sequence to create a spatial-temporal latent representation. This technique highlights several critical considerations: *Temporal dimension variability* – given the varying durations of training videos, the temporal dimension

of the latent space representation cannot be fixed. To address this, one can either sample a specific number of frames (padding or temporal interpolation [34] may be needed for much shorter videos) or define a universally extended (super long) input length for subsequent processing (more details are described in Section 3.2.4); *Utilization of pre-trained visual encoders* – for processing videos of high resolution, leveraging existing pre-trained visual encoders, such as the VAE encoder from Stable Diffusion [19], is advisable for most researchers while Sora's team is expected to train their own compression network with a decoder (the video generator) from scratch via the manner employed in training latent diffusion models [19, 35, 36]. These encoders can efficiently compress large-size patches (e.g., 256×256), facilitating the management of large-scale data; *Temporal information aggregation* – since this method primarily focuses on spatial patch compression, it necessitates an additional mechanism for aggregating temporal information within the model. This aspect is crucial for capturing dynamic changes over time and is further elaborated in subsequent sections (see details in Section 3.3.1 and Figure 14).

Spatial-temporal-patch Compression. This technique is designed to encapsulate both spatial and temporal dimensions of video data, offering a comprehensive representation. This technique extends beyond merely analyzing static frames by considering the movement and changes across frames, thereby capturing the video's dynamic aspects. The utilization of 3D convolution emerges as a straightforward and potent method for achieving this integration [37]. The graphical illustration and the comparison against pure spatial-pachifying are depicted in Figure 9. Similar to spatial-patch compression, employing spatial-temporal-patch compression with predetermined convolution kernel parameters – such as fixed kernel sizes, strides, and output channels – results in variations in the dimensions of the latent space due to the differing characteristics of video inputs. This variability is primarily driven by the diverse durations and resolutions of the videos being processed. To mitigate this challenge, the approaches adopted for spatial patchification are equally applicable and effective in this context.

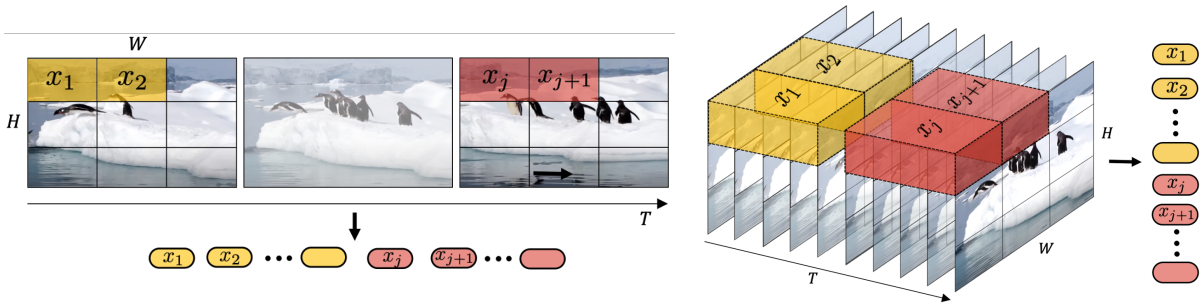
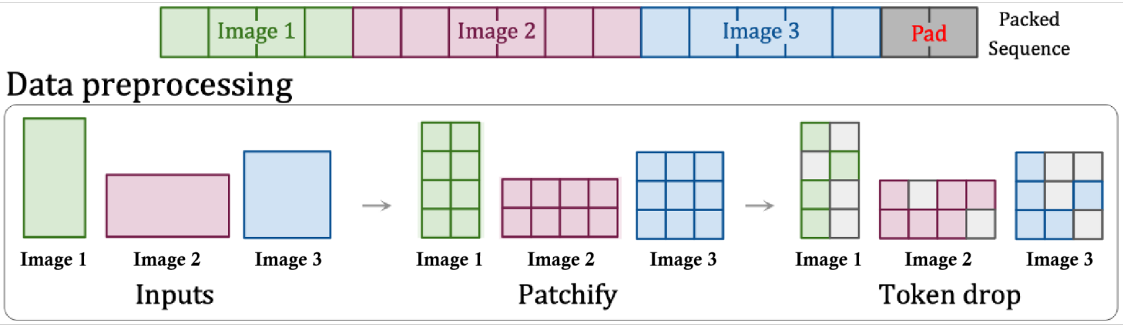


Figure 9: Comparison between different patchification for video compression. Source: ViViT [38]. (**Left**) Spatial patchification simply samples n_t frames and embeds each 2D frame independently following ViT. (**Right**) Spatial-temporal patchification extracts and linearly embeds non-overlapping or overlapping tubelets that span the spatiotemporal input volume.

In summary, we reverse engineer the two patch-level compression approaches based on VAE or its variant like VQ-VQE because operations on patches are more flexible to process different types of videos. Since Sora aims to generate high-fidelity videos, a large patch size or kernel size is used for efficient compression. Here, we expect that fixed-size patches are used for simplicity, scalability, and training stability. But varying-size patches could also be used [39] to make the dimension of the whole frames or videos in latent space consistent. However, it may result in invalid positional encoding, and cause challenges for the decoder to generate videos with varying-size latent patches.



I like this idea already! Neatly embed images into patches sequence almost similar to tokenization of words

Figure 10: Patch packing enables variable resolution images or videos with preserved aspect ratio.6 Token dropping somehow could be treated as data augmentation. Source: NaViT [40].

3.2.4 Spacetime Latent Patches

There is a pivotal concern remaining in the compression network part: How to handle the variability in latent space dimensions (i.e., the number of latent feature chunks or patches from different video types) before feeding patches into the input layers of the diffusion transformer. Here, we discuss several solutions.

Based on Sora's technical report and the corresponding references, **patch n' pack (PNP)** [40] is likely the solution. PNP packs multiple patches from different images in a single sequence as shown in Figure 10. This method is inspired by example packing used in natural language processing [41] that accommodates efficient training on variable length inputs by dropping tokens. Here the patchification and token embedding steps need to be completed in the compression network, but Sora may further patchify the latent for transformer token as Diffusion Transformer does [4]. Regardless there is a second-round patchification or not, we need to address two concerns, how to pack those tokens in a compact manner and how to control which tokens should be dropped. For the first concern, a simple greedy approach is used which adds examples to the first sequence with enough remaining space. Once no more example can fit, sequences are filled with padding tokens, yielding the fixed sequence lengths needed for batched operations. Such a simple packing algorithm can lead to significant padding, depending on the distribution of the length of inputs. On the other hand, we can control the resolutions and frames we sample to ensure efficient packing by tuning the sequence length and limiting padding. For the second concern, an intuitive approach is to drop the similar tokens [42, 43, 33, 44] or, like PNP, apply dropping rate schedulers. However, it is worth noting that ***3D Consistency*** is one of the nice properties of Sora. Dropping tokens may ignore fine-grained details during training. Thus, we believe that OpenAI is likely to use a super long context window and pack all tokens from videos although doing so is computationally expensive e.g., the multi-head attention [45, 46] operator exhibits quadratic cost in sequence length. Specifically, spacetime latent patches from a long-duration video can be packed in one sequence while the ones from several short-duration videos are concatenated in the other sequence.

3.2.5 Discussion

We discuss two technical solutions to data pre-processing that Sora may use. Both solutions are performed at the patch level due to the characteristics of flexibility and scalability for modeling. Different from previous approaches where videos are resized, cropped, or trimmed to a standard size, Sora trains on data at its native size. Although there are several benefits (see detailed analysis in Section 3.2.1), it brings some technical challenges, among which one of the most significant is that neural networks cannot inherently process visual data of variable durations, resolutions, and aspect ratios. Through reverse engineering, we believe that Sora firstly compresses visual patches into low-dimensional latent representations, and arranges such latent patches or further patchified latent patches in a sequence, then injects noise into these latent patches

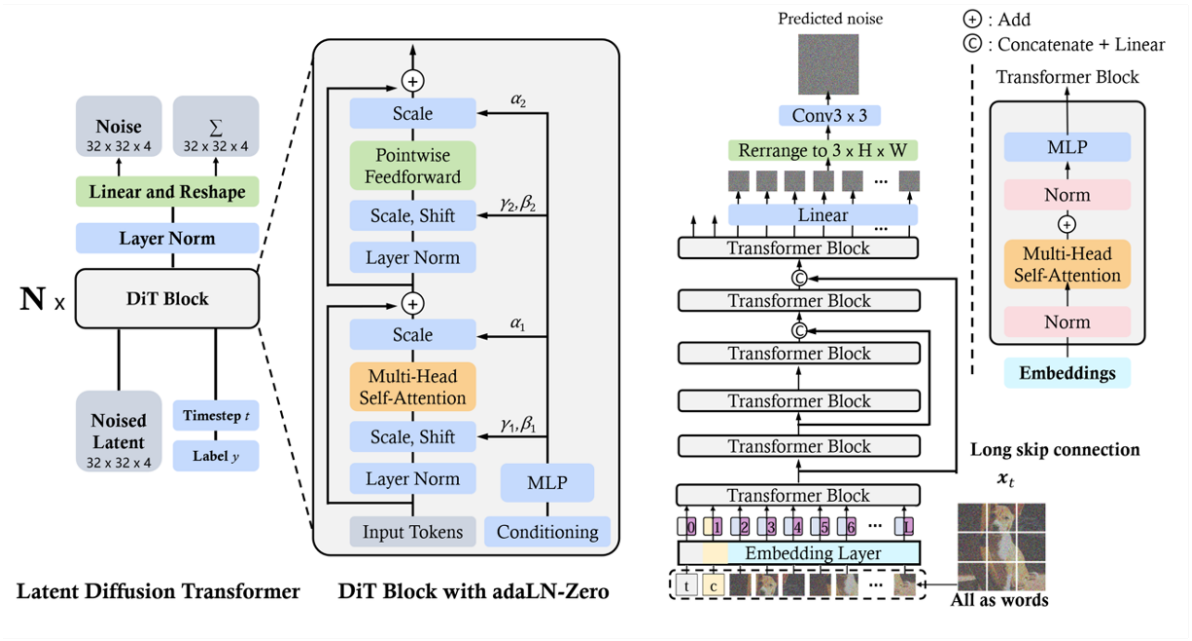


Figure 11: The overall framework of DiT (left) and U-ViT (right)

.

before feeding them to the input layer of diffusion transformer. Spatial-temporal patchification is adopted by `Sora` because it is simple to implement, and it can effectively reduce the context length with high-information-density tokens and decrease the complexity of subsequent modeling of temporal information.

To the research community, we recommend using cost-efficient alternative solutions for video compression and representation, including utilizing pre-trained checkpoints (e.g., compression network) [47], shortening the context window, using light-weight modeling mechanisms such as (grouped) multi-query attention [48, 49] or efficient architectures (e.g. Mamba [50]), downsampling data and dropping tokens if necessary. The trade-off between effectiveness and efficiency for video modeling is an important research topic to be explored.

3.3 Modeling

3.3.1 Diffusion Transformer

Image Diffusion Transformer. Traditional diffusion models [51, 52, 53] mainly leverage convolutional U-Nets that include downsampling and upsampling blocks for the denoising network backbone. However, recent studies show that the U-Net architecture is not crucial to the good performance of the diffusion model. By incorporating a more flexible transformer architecture, the transformer-based diffusion models can use more training data and larger model parameters. Along this line, DiT [4] and U-ViT [54] are among the first works to employ vision transformers for latent diffusion models. As in ViT, DiT employs a multi-head self-attention layer and a pointwise feed-forward network interlaced with some layer norm and scaling layers. Moreover, as shown in Figure 11, DiT incorporates conditioning via adaptive layer norm (AdaLN) with an additional MLP layer for zero-initializing, which initializes each residual block as an identity function and thus greatly stabilizes the training process. The scalability and flexibility of DiT is empirically validated. DiT becomes the new backbone for diffusion models. In U-ViT, as shown in Figure 11, they treat all inputs,

including time, condition, and noisy image patches, as tokens and propose long skip connections between the shallow and deep transformer layers. The results suggest that the downsampling and upsampling operators in CNN-based U-Net are not always necessary, and U-ViT achieves record-breaking FID scores in image and text-to-image generation.

Like Masked AutoEncoder (MAE) [33], Masked Diffusion Transformer (MDT) [55] incorporates mask latent modeling into the diffusion process to explicitly enhance contextual relation learning among object semantic parts in image synthesis. Specifically, as shown in Figure 12, MDT uses a side-interpolated for an additional masked token reconstruction task during training to boost the training efficiency and learn powerful context-aware positional embedding for inference. Compared to DiT [4], MDT achieves better performance and faster learning speed. Instead of using AdaLN (i.e., shifting and scaling) for time-conditioning modeling, Hatamizadeh et al. [56] introduce Diffusion Vision Transformers (DiffiT), which uses a time-dependent self-attention (TMSA) module to model dynamic denoising behavior over sampling time steps. Besides, DiffiT uses two hybrid hierarchical architectures for efficient denoising in the pixel space and the latent space, respectively, and achieves new state-of-the-art results across various generation tasks. Overall, these studies show promising results in employing vision transformers for image latent diffusion, paving the way for future studies for other modalities.

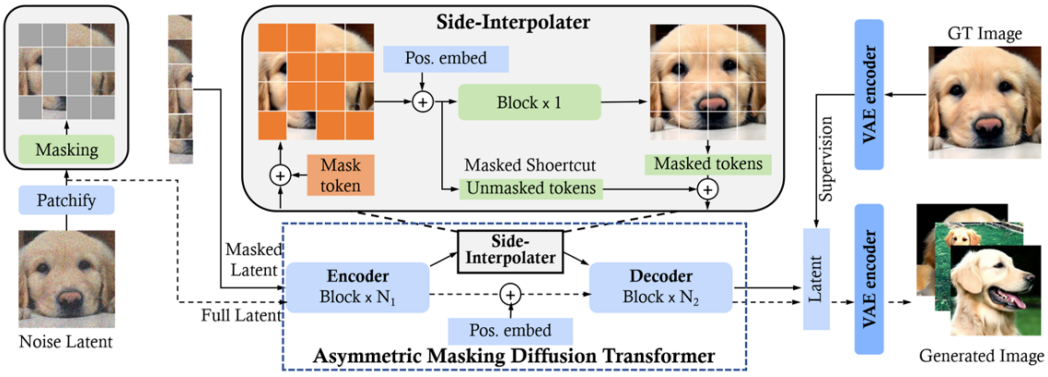
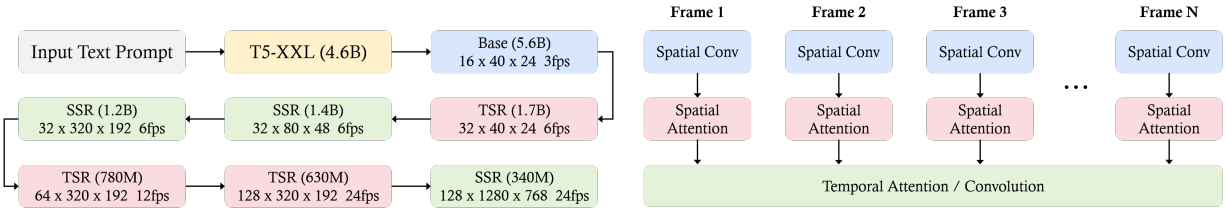


Figure 12: The overall framework of Masked Diffusion Transformer (MDT). A solid/dotted line indicates the training/inference process for each time step. Masking and side-interpolater are only used during training and are removed during inference.

Video Diffusion Transformer. Building upon the foundational works in text-to-image (T2I) diffusion models, recent research has been focused on realizing the potential of diffusion transformers for text-to-video (T2V) generation tasks. Due to the temporal nature of videos, key challenges for applying DiTs in the video domain are: *i) how to compress the video spatially and temporally to a latent space for efficient denoising*; *ii) how to convert the compressed latent to patches and feed them to the transformer;* and *iii) how to handle long-range temporal and spatial dependencies and ensure content consistency*. Please refer to Section 3.2.3 for the first challenge. In this Section, we focus our discussion on transformer-based denoising network architectures designed to operate in the spatially and temporally compressed latent space. We give a detailed review of the two important works (Imagen Video [29] and Video LDM [36]) described in the reference list of the OpenAI `Sora` technique report.

Imagen Video [29], a text-to-video generation system developed by Google Research, utilizes a cascade of diffusion models, which consists of 7 sub-models that perform text-conditional video generation, spatial super-resolution, and temporal super-resolution, to transform textual prompts into high-definition videos. As shown in Figure 13, firstly, a frozen T5 text encoder generates contextual embeddings from the input

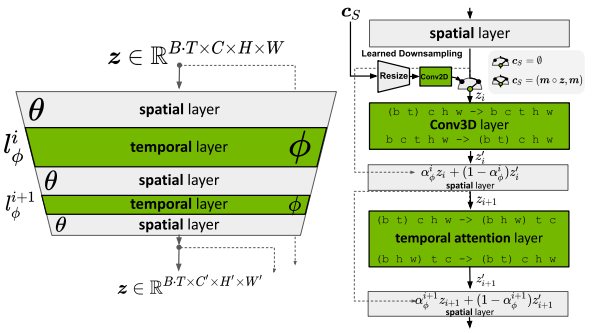


(a) **Cascaded diffusion models.** The cascaded sampling pipeline with a base diffusion model and six up-sampling models that operate spatially and temporally. The text embeddings are injected into all the diffusion models.

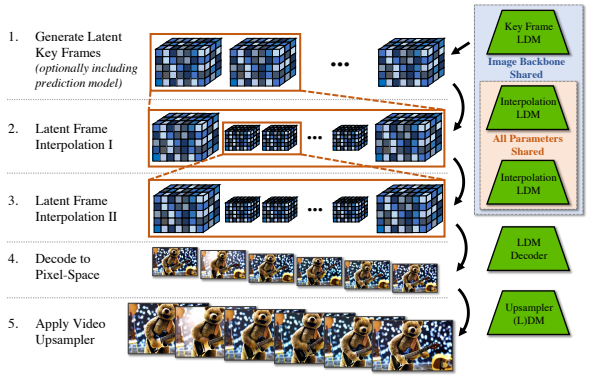
(b) **Video U-Net space-time separable block.** Spatial operations are performed independently over frames with shared parameters, whereas the temporal operation mixes activations over frames. Temporal attention is only used in the base model for memory efficiency.

Figure 13: The overall framework of Imagen Video. Source: Imagen Video [29].

text prompt. These embeddings are critical for aligning the generated video with the text prompt and are injected into all models in the cascade, in addition to the base model. Subsequently, the embedding is fed to the base model for low-resolution video generation, which is then refined by cascaded diffusion models to increase the resolution. The base video and super-resolution models use a 3D U-Net architecture in a space-time separable fashion. This architecture weaves temporal attention and convolution layers with spatial counterparts to efficiently capture inter-frame dependencies. It employs v-prediction parameterization for numerical stability and conditioning augmentation to facilitate parallel training across models. The process involves joint training on both images and videos, treating each image as a frame to leverage larger datasets, and using classifier-free guidance [57] to enhance prompt fidelity. Progressive distillation [58] is applied to streamline the sampling process, significantly reducing the computational load while maintaining perceptual quality. Combining these methods and techniques allows Imagen Video to generate videos with not only high fidelity but also remarkable controllability, as demonstrated by its ability to produce diverse videos, text animations, and content in various artistic styles.



(a) **Additional temporal layer**. A pre-trained LDM is turned into a video generator by inserting temporal layers that learn to align frames into temporally consistent sequences. During optimization, the image backbone θ remains fixed and only the parameters ϕ of the temporal layers l^i_ϕ are trained.



(b) **Video LDM stack.** Video LDM first generates sparse key frames and then temporally interpolates twice with the same latent diffusion models to achieve a high frame rate. Finally, the latent video is decoded to pixel space, and optionally, a video upsampler diffusion model is applied.

Figure 14: The overall framework of Video LDM. Source: Video LDM [36].

Blattmann et al. [36] propose to turn a 2D Latent Diffusion Model into a Video Latent Diffusion Model (Video LDM). They achieve this by adding some post-hoc temporal layers among the existing spatial layers into both the U-Net backbone and the VAE decoder that learns to align individual frames. These temporal layers are trained on encoded video data, while the spatial layers remain fixed, allowing the model to

leverage large image datasets for pre-training. The LDM's decoder is fine-tuned for temporal consistency in pixel space and temporally aligning diffusion model upsamplers for enhanced spatial resolution. To generate very long videos, models are trained to predict a future frame given a number of context frames, allowing for classifier-free guidance during sampling. To achieve high temporal resolution, the video synthesis process is divided into key frame generation and interpolation between these key frames. Following cascaded LDMs, a DM is used to further scale up the Video LDM outputs by 4 times, ensuring high spatial resolution while maintaining temporal consistency. This approach enables the generation of globally coherent long videos in a computationally efficient manner. Additionally, the authors demonstrate the ability to transform pre-trained image LDMs (e.g., Stable Diffusion) into text-to-video models by training only the temporal alignment layers, achieving video synthesis with resolutions up to 1280×2048.

3.3.2 Discussion

Cascade diffusion models for spatial and temporal up-sampling. `Sora` can generate high-resolution videos. By reviewing existing works and our reverse engineering, we speculate that `Sora` also leverages cascade diffusion model architecture [59] which is composed of a base model and many space-time refiner models. The attention modules are unlikely to be heavily used in the based diffusion model and low-resolution diffusion model, considering the high computation cost and limited performance gain of using attention machines in high-resolution cases. For spatial and temporal scene consistency, as previous works show that temporal consistency is more important than spatial consistency for video/scene generation, `Sora` is likely to leverage an efficient training strategy by using longer video (for temporal consistency) with lower resolution. Moreover, `Sora` is likely to use a v-parameterization diffusion model [58], considering its superior performance compared to other variants that predict the original latent x or the noise ϵ.

On the latent encoder. For training efficiency, most of the existing works leverage the pre-trained VAE encoder of Stable Diffusions [60, 61], a pre-trained 2D diffusion model, as an initialized model checkpoint. However, the encoder lacks the temporal compression ability. Even though some works propose to only fine-tune the decoder for handling temporal information, the decoder's performance of dealing with video temporal data in the compressed latent space remains sub-optimal. Based on the technique report, our reverse engineering shows that, instead of using an existing pre-trained VAE encoder, it is likely that `Sora` uses a space-time VAE encoder, trained from scratch on video data, which performs better than existing ones with a video-orient compressed latent space.

3.4 Language Instruction Following

Users primarily engage with generative AI models through natural language instructions, known as text prompts [62, 63]. Model instruction tuning aims to enhance AI models' capability to follow prompts accurately. This improved capability in prompt following enables models to generate output that more closely resembles human responses to natural language queries. We start our discussion with a review of instruction following techniques for large language models (LLMs) and text-to-image models such as DALL·E 3. To enhance the text-to-video model's ability to follow text instructions, `Sora` utilizes an approach similar to that of DALL·E 3. The approach involves training a descriptive captioner and utilizing the captioner's generated data for fine-tuning. As a result of instruction tuning, `Sora` is able to accommodate a wide range of user requests, ensuring meticulous attention to the details in the instructions and generating videos that precisely meet users' needs.

3.4.1 Large Language Models

The capability of LLMs to follow instructions has been extensively explored [64, 65, 66]. This ability allows LLMs to read, understand, and respond appropriately to instructions describing an unseen task without examples. Prompt following ability is obtained and enhanced by fine-tuning LLMs on a mixture of tasks

formatted as instructions[64, 66], known as instruction tuning. Wei et al. [65] showed that instruction-tuned LLMs significantly outperform the untuned ones on unseen tasks. The instruction-following ability transforms LLMs into general-purpose task solvers, marking a paradigm shift in the history of AI development.

3.4.2 Text-to-Image

The instruction following in DALL·E 3 is addressed by a caption improvement method with a hypothesis that the quality of text-image pairs that the model is trained on determines the performance of the resultant text-to-image model [67]. The poor quality of data, particularly the prevalence of noisy data and short captions that omit a large amount of visual information, leads to many issues such as neglecting keywords and word order, and misunderstanding the user intentions [21]. The caption improvement approach addresses these issues by re-captioning existing images with detailed, descriptive captions. The approach first trains an image captioner, which is a vision-language model, to generate precise and descriptive image captions. The resulting descriptive image captions by the captioner are then used to fine-tune text-to-image models. Specifically, DALL·E 3 follows contrastive captioners (CoCa) [68] to jointly train an image captioner with a CLIP [26] architecture and a language model objective. This image captioner incorporates an image encoder a unimodal text encoder for extracting language information, and a multimodal text decoder. It first employs a contrastive loss between unimodal image and text embeddings, followed by a captioning loss for the multimodal decoder's outputs. The resulting image captioner is further fine-tuned on a highly detailed description of images covering main objects, surroundings, backgrounds, texts, styles, and colorations. With this step, the image captioner is able to generate detailed descriptive captions for the images. The training dataset for the text-to-image model is a mixture of the re-captioned dataset generated by the image captioner and ground-truth human-written data to ensure that the model captures user inputs. This image caption improvement method introduces a potential issue: a mismatch between the actual user prompts and descriptive image descriptions from the training data. DALL·E 3 addresses this by *upsampling*, where LLMs are used to re-write short user prompts into detailed and lengthy instructions. This ensures that the model's text inputs received in inference time are consistent with those in model training.

3.4.3 Text-to-Video

To enhance the ability of instruction following, `Sora` adopts a similar caption improvement approach. This method is achieved by first training a video captioner capable of producing detailed descriptions for videos. Then, this video captioner is applied to all videos in the training data to generate high-quality (video, descriptive caption) pairs, which are used to fine-tune `Sora` to improve its instruction following ability.

`Sora`'s technical report [3] does not reveal the details about how the video captioner is trained. Given that the video captioner is a video-to-text model, there are many approaches to building it. A straightforward approach is to utilize CoCa architecture for video captioning by taking multiple frames of a video and feeding each frame into the image encoder [68], known as VideoCoCa [69]. VideoCoCa builds upon CoCa and re-uses the image encoder pre-trained weights and applies it independently on sampled video frames. The resulting frame token embeddings are flattened and concatenated into a long sequence of video representations. These flattened frame tokens are then processed by a generative pooler and a contrastive pooler, which are jointly trained with the contrastive loss and captioning loss. Other alternatives to building video captioners include mPLUG-2 [70], GIT [71], FrozenBiLM [72], and more. Finally, to ensure that user prompts align with the format of those descriptive captions in training data, `Sora` performs an additional prompt extension step, where GPT-4V is used to expand user inputs to detailed descriptive prompts.

3.4.4 Discussion

The instruction-following ability is critical for Sora to generate one-minute-long videos with intricate scenes that are faithful to user intents. According to Sora's technical report [3], this ability is obtained by developing a captioner that can generate long and detailed captions, which are then used to train the model. However, the process of collecting data for training such a captioner is unknown and likely labor-intensive, as it may require detailed descriptions of videos. Moreover, the descriptive video captioner might hallucinate important details of the videos. We believe that how to improve the video captioner warrants further investigation and is critical to enhance the instruction-following ability of text-to-image models.

3.5 Prompt Engineering

Prompt engineering refers to the process of designing and refining the input given to an AI system, particularly in the context of generative models, to achieve specific or optimized outputs [73, 74, 75]. The art and science of prompt engineering involve crafting these inputs in a way that guides the model to produce the most accurate, relevant, and coherent responses possible.

3.5.1 Text Prompt

Text prompt engineering is vital in directing text-to-video models (e.g., Sora [3]) to produce videos that are visually striking while precisely meeting user specifications. This involves crafting detailed descriptions to instruct the model to effectively bridge the gap between human creativity and AI's execution capabilities [76]. The prompts for Sora cover a wide range of scenarios. Recent works (e.g., VoP [77], Make-A-Video [28], and Tune-A-Video [78]) have shown how prompt engineering leverages model's natural language understanding ability to decode complex instructions and render them into cohesive, lively, and high-quality video narratives. As shown in Figure 15, "a stylish woman walking down a neon-lit Tokyo street..." is such a meticulously crafted text prompt that it ensures Sora to generate a video that aligns well with the expected vision. The quality of prompt engineering depends on the careful selection of words, the specificity of the details provided, and comprehension of their impact on the model's output. For example, the prompt in Figure 15 specifies in detail the actions, settings, character appearances, and even the desired mood and atmosphere of the scene.

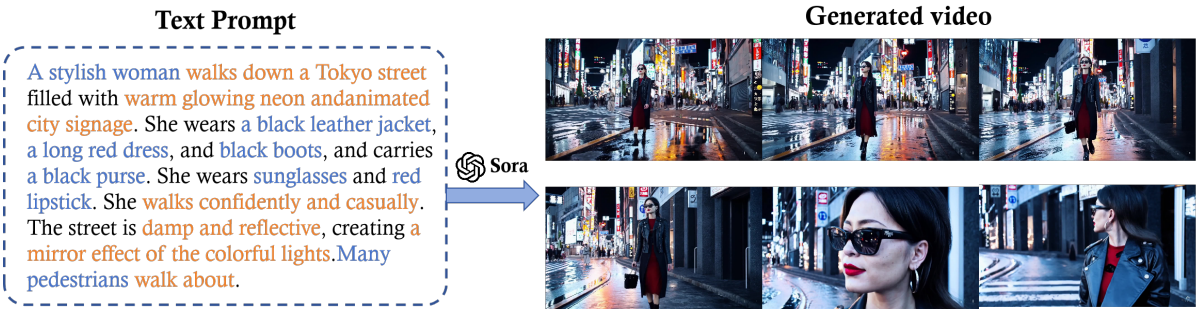


Figure 15: A case study on prompt engineering for text-to-video generation, employing color coding to delineate the creative process. The text highlighted in blue describes the elements generated by Sora, such as the depiction of a stylish woman. In contrast, the text in yellow accentuates the model's interpretation of actions, settings, and character appearances, demonstrating how a meticulously crafted prompt is transformed into a vivid and dynamic video narrative.

3.5.2 Image Prompt

An image prompt serves as a visual anchor for the to-be-generated video's content and other elements such as characters, setting, and mood [79]. In addition, a text prompt can instruct the model to animate these elements by e.g., adding layers of movement, interaction, and narrative progression that bring the static image to life [27, 80, 81]. The use of image prompts allows Sora to convert static images into dynamic, narrative-driven videos by leveraging both visual and textual information. In Figure 16, we show AI-generated videos of "a Shiba Inu wearing a beret and turtleneck", "a unique monster family", "a cloud forming the word 'SORA"' and "surfers navigating a tidal wave inside a historic hall". These examples demonstrate what can be achieved by prompting Sora with DALL·E-generated images.

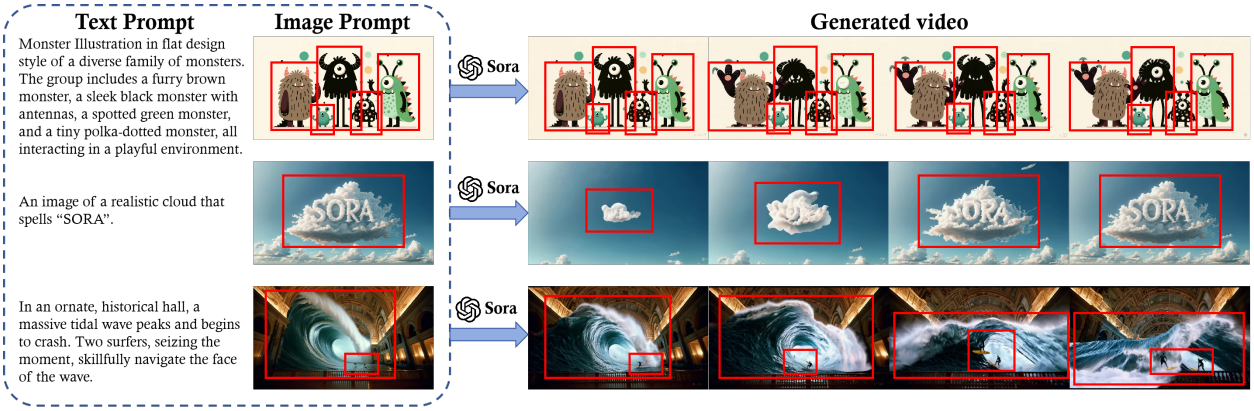


Figure 16: This example illustrates the image prompts to guide Sora's text-to-video model to generation. The red boxes visually anchor the key elements of each scene—monsters of varied designs, a cloud formation spelling "SORA", and surfers in an ornate hall facing a massive tidal wave.

3.5.3 Video Prompt

Video prompts can also be used for video generation as demonstrated in [82, 83]. Recent works (e.g., Moonshot [84] and Fast-Vid2Vid [85]) show that good video prompts need to be specific and flexible. This ensures that the model receives clear direction on specific objectives, like the portrayal of particular objects and visual themes, and also allows for imaginative variations in the final output. For example, in the video extension tasks, a prompt could specify the direction (forward or backward in time) and the context or theme of the extension. In Figure 17(a), the video prompt instructs Sora to extend a video backward in time to explore the events leading up to the original starting point. When performing video-to-video editing through video prompts, as shown in Figure 17(b), the model needs to clearly understand the desired transformation, such as changing the video's style, setting or atmosphere, or altering subtle aspects like lighting or mood. In Figure 17(c), the prompt instructs Sora to connect videos while ensuring smooth transitions between objects in different scenes across videos.

3.5.4 Discussion

Prompt engineering allows users to guide AI models to generate content that aligns with their intent. As an example, the combined use of text, image, and video prompts enables Sora to create content that is not only visually compelling but also aligned well with users' expectations and intent. While previous studies on prompt engineering have been focused on text and image prompts for LLMs and LVMs [86, 87, 88], we expect that there will be a growing interest in video prompts for video generation models.

3.6 Trustworthiness

With the rapid advancement of sophisticated models such as ChatGPT [89], GPT4-V [90], and Sora [3], the capabilities of these models have seen remarkable enhancements. These developments have made sig-

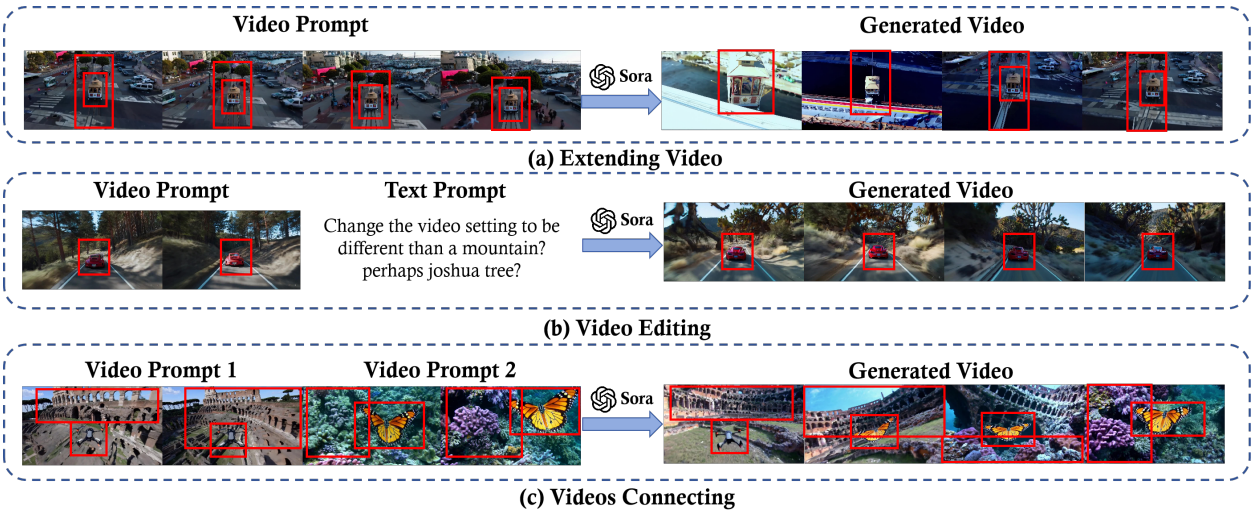


Figure 17: These examples illustrate the video prompt techniques for Sora models: (a) Video Extension, where the model extrapolates the sequence backward the original footage, (b) Video Editing, where specific elements like the setting are transformed as per the text prompt, and (c) Video Connection, where two distinct video prompts are seamlessly blended to create a coherent narrative. Each process is guided by a visual anchor, marked by a red box, ensuring continuity and precision in the generated video content.

nificant contributions to improving work efficiency and propelling technological progress. However, these advancements also raise concerns about the potential for misuse of these technologies, including the generation of fake news [91, 92], privacy breaches [93], and ethical dilemmas [94, 95]. Consequently, the issue of trustworthiness in large models has garnered extensive attention from both the academic and industrial spheres, emerging as a focal point of contemporary research discussions.

3.6.1 Safety Concern

One primary area of focus is the model's safety, specifically its resilience against misuse and so-called "jailbreak" attacks, where users attempt to exploit vulnerabilities to generate prohibited or harmful content [96, 97, 98, 99, 100, 101, 102, 103, 104, 105]. For instance, AutoDAN [103], a novel and interpretable adversarial attack method based on gradient techniques, is introduced to enable system bypass. In a recent study, researchers explore two reasons why LLMs struggle to resist jailbreak attacks: competing objectives and mismatched generalization [106]. Besides textual attacks, visual jailbreak also threatens the safety of multimodal models (e.g., GPT-4V [90], and Sora [3]). A recent study [107] found that large multimodal models are more vulnerable since the continuous and high-dimensional nature of the additional visual input makes it weaker against adversarial attacks, representing an expanded attack surface.

3.6.2 Other Exploitation

Due to the large scale of the training dataset and training methodology of large foundation models (e.g., ChatGPT [89] and Sora [3]), the truthfulness of these models needs to be enhanced as the related issues like hallucination have been discussed widely [108]. Hallucination in this context refers to the models' tendency to generate responses that may appear convincing but are unfounded or false [96]. This phenomenon raises critical questions about the reliability and trustworthiness of model outputs, necessitating a comprehensive approach to both evaluate and address the issue. Amount of studies have been dedicated to dissecting the problem of hallucination from various angles. This includes efforts aimed at evaluating the extent and nature of hallucination across different models and scenarios [109, 96, 110, 111]. These evaluations provide invaluable insights into how and why hallucinations occur, laying the groundwork for

developing strategies to mitigate their incidence. Concurrently, a significant body of research is focused on devising and implementing methods to reduce hallucinations in these large models [112, 113, 114].

Another vital aspect of trustworthiness is fairness and bias. The critical importance of developing models that do not perpetuate or exacerbate societal biases is a paramount concern. This priority stems from the recognition that biases encoded within these models can reinforce existing social inequities, leading to discriminatory outcomes. Studies in this area, as evidenced by the work of Gallegos et al. [115], Zhang et al. [116], Liang et al. [117], and Friedrich et al. [118], are dedicated to the meticulous identification and rectification of these inherent biases. The goal is to cultivate models that operate fairly, treating all individuals equitably without bias towards race, gender, or other sensitive attributes. This involves not only detecting and mitigating bias in datasets but also designing algorithms that can actively counteract the propagation of such biases [119, 120].

Privacy preservation emerges as another foundational pillar when these models are deployed. In an era where data privacy concerns are escalating, the emphasis on protecting user data has never been more critical. The increasing public awareness and concern over how personal data is handled have prompted more rigorous evaluations of large models. These evaluations focus on the models' capacity to protect user data, ensuring that personal information remains confidential and is not inadvertently disclosed. Research by Mireshghallah et al. [121], Plant et al. [122], and Li et al. [123] exemplify efforts to advance methodologies and technologies that safeguard privacy.

3.6.3 Alignment

In addressing these challenges, ensuring the trustworthiness of large models has become one of the primary concerns for researchers [124, 96, 99, 125]. Among the most important technologies is model alignment [125, 126], which refers to the process and goal of ensuring that the behavior and outputs of models are consistent with the intentions and ethical standards of human designers. This concerns the development of technology, its moral responsibilities, and social values. In the domain of LLMs, the method of Reinforcement Learning with Human Feedback (RLHF) [127, 128] has been widely applied for model alignment. This method combines Reinforcement Learning (RL) with direct human feedback, allowing models to better align with human expectations and standards in understanding and performing tasks.

3.6.4 Discussion

From `Sora` (specifically its technical report), we summarize some insightful findings that potentially offer an informative guideline for future work:

(1) *Integrated Protection of Model and External Security*: As models become more powerful, especially in generating content, ensuring that they are not misused to produce harmful content (such as hate speech [129] and false information [92, 91]) has become a serious challenge. In addition to aligning the model itself, external security protections are equally important. This includes content filtering and review mechanisms, usage permissions and access control, data privacy protection, as well as enhancements in transparency and explainability. For instance, OpenAI now uses a detection classifier to tell whether a given video is generated by `Sora` [130]. Moreover, a text classifier is deployed to detect the potentially harmful textual input [130].

(2) *Security Challenges of Multimodal Models*: Multimodal models, such as text-to-video models like `Sora` bring additional complexity to security due to their ability to understand and generate various types of content (text, images, videos, etc.). Multimodal models can produce content in various forms, increasing the ways and scope of misuse and copyright issues. As the content generated by multimodal models is more complex and diverse, traditional methods of content verification and authenticity may no longer be effective. This requires the development of new technologies and methods to identify and filter harmful content generated by these models, increasing the difficulty of regulation and management.

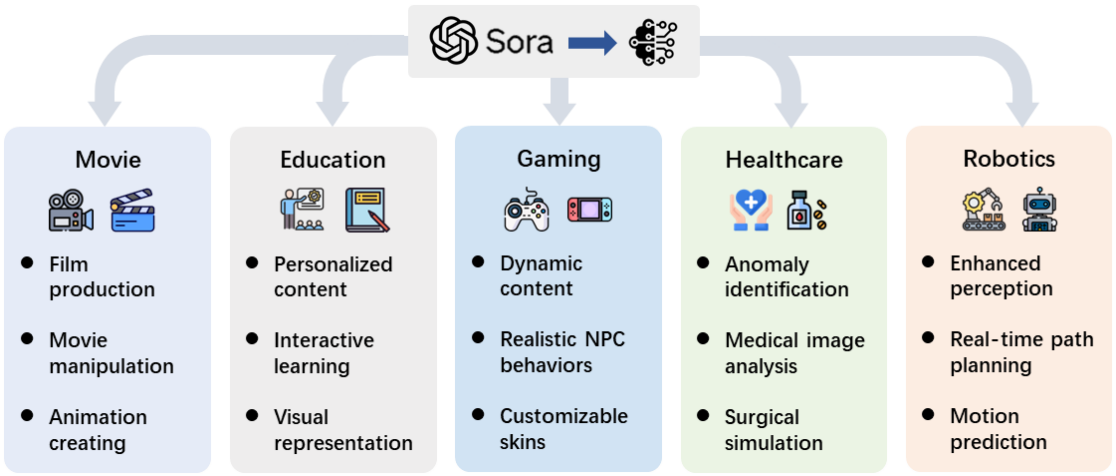


Figure 18: Applications of Sora.

(3) *The Need for Interdisciplinary Collaboration*: Ensuring the safety of models is not just a technical issue but also requires cross-disciplinary cooperation. To address these challenges, experts from various fields such as law [131] and psychology [132] need to work together to develop appropriate norms (e.g., what's the safety and what's unsafe?), policies, and technological solutions. The need for interdisciplinary collaboration significantly increases the complexity of solving these issues.

4 Applications

As video diffusion models, exemplified by Sora, emerge as a forefront technology, their adoption across diverse research fields and industries is rapidly accelerating. The implications of this technology extend far beyond mere video creation, offering transformative potential for tasks ranging from automated content generation to complex decision-making processes. In this section, we delve into a comprehensive examination of the current applications of video diffusion models, highlighting key areas where Sora has not only demonstrated its capabilities but also revolutionized the approach to solving complex problems. We aim to offer a broad perspective for the practical deployment scenarios (see Figure 18).

4.1 Movie

Traditionally, creating cinematic masterpieces has been an arduous and expensive process, often requiring decades of effort, cutting-edge equipment, and substantial financial investments. However, the advent of advanced video generation technologies heralds a new era in film-making, one where the dream of autonomously producing movies from simple text inputs is becoming a reality. Researchers have ventured into the realm of movie generation by extending video generation models into creating movies. MovieFactory [133] applies diffusion models to generate film-style videos from elaborate scripts produced by ChatGPT [89], representing a significant leap forward. In the follow-up, MobileVidFactory [134] can automatically generate vertical mobile videos with only simple texts provided by users. Vlogger [135] makes it feasible for users to compose a minute-long vlog. These developments, epitomized by Sora's ability to generate captivating movie content effortlessly, mark a pivotal moment in the democratization of movie production. They offer a glimpse into a future where anyone can be a filmmaker, significantly lowering the barriers to entry in the film industry and introducing a novel dimension to movie production that blends traditional storytelling with AI-driven creativity. The implications of these technologies extend beyond simplification. They promise to reshape the landscape of film production, making it more accessible and

versatile in the face of evolving viewer preferences and distribution channels.

4.2 Education

The landscape of educational content has long been dominated by static resources, which, despite their value, often fall short of catering to the diverse needs and learning styles of today's students. Video diffusion models stand at the forefront of an educational revolution, offering unprecedented opportunities to customize and animate educational materials in ways that significantly enhance learner engagement and understanding. These advanced technologies enable educators to transform text descriptions or curriculum outlines into dynamic, engaging video content tailored to the specific style, and interests of individual learners [136, 137, 138, 139]. Moreover, image-to-video editing techniques [140, 141, 142] present innovative avenues for converting static educational assets into interactive videos, thereby supporting a range of learning preferences and potentially increasing student engagement. By integrating these models into educational content creation, educators can produce videos on a myriad of subjects, making complex concepts more accessible and captivating for students. The use of `Sora` in revolutionizing the educational domain exemplifies the transformative potential of these technologies. This shift towards personalized, dynamic educational content heralds a new era in education.

4.3 Gaming

The gaming industry constantly seeks ways to push the boundaries of realism and immersion, yet traditional game development often grapples with the limitations of pre-rendered environments and scripted events. The generation of dynamic, high-fidelity video content and realistic sound by diffusion models effects in real-time, promise to overcome existing constraints, offering developers the tools to create evolving game environments that respond organically to player actions and game events [143, 144]. This could include generating changing weather conditions, transforming landscapes, or even creating entirely new settings on the fly, making game worlds more immersive and responsive. Some methods [145, 146] also synthesize realistic impact sounds from video inputs, enhancing game audio experiences. With the integration of `Sora` within the gaming domain, unparalleled immersive experiences that captivate and engage players can be created. How games are developed, played, and experienced will be innovated, as well as opening new possibilities for storytelling, interaction, and immersion.

4.4 Healthcare

Despite generative capabilities, video diffusion models excel in understanding and generating complex video sequences, making them particularly suited for identifying dynamic anomalies within the body, such as early cellular apoptosis [147], skin lesion progression [148], and irregular human movements [149], which are crucial for early disease detection and intervention strategies. Additionally, models like MedSegDiff-V2 [150] and [151] leverage the power of transformers to segment medical images with unprecedented precision, enabling clinicians to pinpoint areas of interest across various imaging modalities with enhanced accuracy. The integration of `Sora` into clinical practice promises not only to refine diagnostic processes but also to personalize patient care, offering tailored treatment plans based on precise medical imaging analysis. However, this technological integration comes with its own set of challenges, including the need for robust data privacy measures and addressing ethical considerations in healthcare.

4.5 Robotics

Video diffusion models now play important roles in robotics, showing a new era where robots can generate and interpret complex video sequences for enhanced perception [152, 153] and decision-making [154, 155, 156]. These models unlock new capabilities in robots, enabling them to interact with their environment and execute tasks with unprecedented complexity and precision. The introduction of web-scale diffusion models to robotics [152] showcases the potential for leveraging large-scale models to enhance robotic vision

and understanding. Latent diffusion models are employed for language-instructed video prediction [157], allowing robots to understand and execute tasks by predicting the outcome of actions in video format. Furthermore, the reliance on simulated environments for robotics research has been innovatively addressed by video diffusion models capable of creating highly realistic video sequences [158, 159]. This enables the generation of diverse training scenarios for robots, mitigating the limitations imposed by the scarcity of real-world data. We believe, the integration of technologies like Sora into the robotics field holds the promise of groundbreaking developments. By harnessing the power of Sora, the future of robotics is poised for unprecedented advancements, where robots can seamlessly navigate and interact with their environments.

5 Discussion

Sora shows a remarkable talent for precisely understanding and implementing complex instructions from humans. This model excels at creating detailed videos with various characters, all set within elaborately crafted settings. A particularly impressive attribute of Sora is its ability to produce videos up to one minute in length while ensuring consistent and engaging storytelling. This marks a significant improvement over previous attempts that focused on shorter video pieces, as Sora's extended sequences exhibit a clear narrative flow and maintain visual consistency from start to finish. Furthermore, Sora distinguishes itself by generating longer video sequences that capture complex movements and interactions, advancing past the restrictions of earlier models that could only handle short clips and basic images. This advancement signifies a major step forward in AI-powered creative tools, enabling users to transform written stories into vivid videos with a level of detail and sophistication that was previously unattainable.

5.1 Limitations

Challenges in Physical Realism. Sora, as a simulation platform, exhibits a range of limitations that undermine its effectiveness in accurately depicting complex scenarios. Most important is its inconsistent handling of physical principles within complex scenes, leading to a failure in accurately copying specific examples of cause and effect. For instance, consuming a portion of a cookie might not result in a corresponding bite mark, illustrating the system's occasional departure from physical plausibility. This issue extends to the simulation of motion, where Sora generates movements that challenge realistic physical modeling, such as unnatural transformations of objects or the incorrect simulation of rigid structures like chairs, leading to unrealistic physical interactions. The challenge further increases when simulating complex interactions among objects and characters, occasionally producing outcomes that lean towards the humorous.

Spatial and Temporal Complexities. Sora occasionally misunderstands instructions related to the placement or arrangement of objects and characters within a given prompt, leading to confusion about directions (e.g., confusing left for right). Additionally, it faces challenges in maintaining the temporal accuracy of events, particularly when it comes to adhering to designated camera movements or sequences. This can result in deviating from the intended temporal flow of scenes. In complex scenarios that involve a multitude of characters or elements, Sora has a tendency to insert irrelevant animals or people. Such additions can significantly change the originally envisioned composition and atmosphere of the scene, moving away from the planned narrative or visual layout. This issue not only affects the model's ability to accurately recreate specific scenes or narratives but also impacts its reliability in generating content that closely aligns with the user's expectations and the coherence of the generated output.

Limitations in Human-computer Interaction (HCI). Sora, while showing potential in the video generation domain, faces significant limitations in HCI. These limitations are primarily evident in the coherence and efficiency of user-system interactions, especially when making detailed modifications or optimizations to generated content. For instance, users might find it difficult to precisely specify or adjust the presentation of specific elements within a video, such as action details and scene transitions. Additionally, Sora's limi-

tations in understanding complex language instructions or capturing subtle semantic differences could result in video content that does not fully meet user expectations or needs. These shortcomings restrict Sora's potential in video editing and enhancement, also impacting the overall satisfaction of the user experience.

Usage Limitation. Regarding usage limitations, OpenAI has not yet set a specific release date for public access to Sora, emphasizing a cautious approach towards safety and readiness before broad deployment. This indicates that further improvements and testing in areas such as security, privacy protection, and content review may still be necessary for Sora. Moreover, at present, Sora can only generate videos up to one minute in length, and according to published cases, most generated videos are only a few dozen seconds long. This limitation restricts its use in applications requiring longer content display, such as detailed instructional videos or in-depth storytelling. This limitation reduces Sora's flexibility in the content creation.

5.2 Opportunities

Academy. (1) The introduction of Sora by OpenAI marks a strategic shift towards encouraging the broader AI community to delve deeper into the exploration of text-to-video models, leveraging both diffusion and transformer technologies. This initiative aims to redirect the focus toward the potential of creating highly sophisticated and nuanced video content directly from textual descriptions, a frontier that promises to revolutionize content creation, storytelling, and information sharing. (2) The innovative approach of training Sora on data at its native size, as opposed to the traditional methods of resizing or cropping, serves as a groundbreaking inspiration for the academic community. It opens up new pathways by highlighting the benefits of utilizing unmodified datasets, which leads to the creation of more advanced generative models.

Industry. (1) The current capabilities of Sora signal a promising path for the advancement of video simulation technologies, highlighting the potential to significantly enhance realism within both physical and digital areas. The prospect of Sora enabling the creation of highly realistic environments through textual descriptions presents a promising future for content creation. This potential extends to revolutionizing game development, offering a glimpse into a future where immersive-generated worlds can be crafted with unprecedented ease and accuracy. (2) Companies may leverage Sora to produce advertising videos that swiftly adapt to market changes and create customized marketing content. This not only reduces production costs but also enhances the appeal and effectiveness of advertisements. The ability of Sora to generate highly realistic video content from textual descriptions alone could revolutionize how brands engage with their audience, allowing for the creation of immersive and compelling videos that capture the essence of their products or services in unprecedented ways.

Society. (1) While the prospect of utilizing text-to-video technology to replace traditional filmmaking remains distant, Sora and similar platforms hold transformative potential for content creation on social media. The constraints of current video lengths do not diminish the impact these tools can have in making high-quality video production accessible to everyone, enabling individuals to produce compelling content without the need for expensive equipment. It represents a significant shift towards empowering content creators across platforms like TikTok and Reels, bringing in a new age of creativity and engagement. (2) Screenwriters and creative professionals can use Sora to transform written scripts into videos, assisting them in better showing and sharing their creative concepts, and even in producing short films and animations. The ability to create detailed, vivid videos from scripts can fundamentally change the pre-production process of filmmaking and animation, offering a glimpse into how future storytellers might pitch, develop, and refine their narratives. This technology opens up possibilities for a more dynamic and interactive form of script development, where ideas can be visualized and assessed in real time, providing a powerful tool for creativity and collaboration. (3) Journalists and news organizations can also utilize Sora to quickly generate news reports or explanatory videos, making the news content more vivid and engaging. This can significantly increase the coverage and audience engagement of news reports. By providing a tool that

can simulate realistic environments and scenarios, Sora offers a powerful solution for visual storytelling, enabling journalists to convey complex stories through engaging videos that were previously difficult or expensive to produce. In summary, Sora's potential to revolutionize content creation across marketing, journalism, and entertainment is immense.

6 Conclusion

We present a comprehensive review of Sora to help developers and researchers study the capabilities and related works of Sora. The review is based on our survey of published technical reports and reverse engineering based on existing literature. We will continue to update the paper when Sora's API is available and further details about Sora are revealed. We hope that this review paper will prove a valuable resource for the open-source research community and lay a foundation for the community to jointly develop an open-source version of Sora in the near future to democratize video auto-creation in the era of AIGC. To achieve this goal, we invite discussions, suggestions, and collaborations on all fronts.

References

[1] OpenAI, "Chatgpt: Get instant answers, find creative inspiration, learn something new.." https://openai.com/chatgpt, 2022.

[2] OpenAI, "Gpt-4 technical report," 2023.

[3] OpenAI, "Sora: Creating video from text." https://openai.com/sora, 2024.

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

[5] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, pp. 1033–1038, IEEE, 1999.

[6] P. S. Heckbert, "Survey of texture mapping," *IEEE computer graphics and applications*, vol. 6, no. 11, pp. 56–67, 1986.

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv*, 2014.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[9] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.

[10] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[11] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," *arXiv preprint arXiv:2303.04226*, 2023.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[17] 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: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[19] 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*, pp. 10684–10695, 2022.

[20] M. AI, "Midjourney: Text to image with ai art generator." https://www.midjourneyai.ai/en, 2023.

[21] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, *et al.*, "Improving image generation with better captions," *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, vol. 2, p. 3, 2023.

[22] P. AI, "Pika is the idea-to-video platform that sets your creativity in motion.." https://pika.art/home, 2023.

[23] R. AI, "Gen-2: Gen-2: The next step forward for generative ai." https://research.runwayml.com/gen2, 2023.

[24] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

[25] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. P. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, *et al.*, "Scaling vision transformers to 22 billion parameters," in *International Conference on Machine Learning*, pp. 7480–7512, PMLR, 2023.

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

[27] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[28] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman, "Make-a-video: Text-to-video generation without text-video data," 2022.

[29] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[30] R. Sutton, "The bitter lesson." http://www.incompleteideas.net/IncIdeas/BitterLesson.html, March 2019. Accessed: Your Access Date Here.

[31] S. Xie, "Take on sora technical report." https://twitter.com/sainingxie/status/1758433676105310543, 2024.

[32] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[33] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

[34] S. Ge, S. Nah, G. Liu, T. Poon, A. Tao, B. Catanzaro, D. Jacobs, J.-B. Huang, M.-Y. Liu, and Y. Balaji, "Preserve your own correlation: A noise prior for video diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22930–22941, 2023.

[35] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," *arXiv preprint arXiv:2311.17042*, 2023.

[36] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023.

[37] M. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, and A. Angelova, "Tokenlearner: Adaptive space-time tokenization for videos," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12786–12797, 2021.

[38] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," *arXiv preprint arXiv:2103.15691*, 2021.

[39] L. Beyer, P. Izmailov, A. Kolesnikov, M. Caron, S. Kornblith, X. Zhai, M. Minderer, M. Tschannen, I. Alabdulmohsin, and F. Pavetic, "Flexivit: One model for all patch sizes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14496–14506, 2023.

[40] M. Dehghani, B. Mustafa, J. Djolonga, J. Heek, M. Minderer, M. Caron, A. Steiner, J. Puigcerver, R. Geirhos, I. M. Alabdulmohsin, *et al.*, "Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[41] M. M. Krell, M. Kosec, S. P. Perez, and A. Fitzgibbon, "Efficient sequence packing without cross-contamination: Accelerating large language models without impacting performance," *arXiv preprint arXiv:2107.02027*, 2021.

[42] H. Yin, A. Vahdat, J. M. Alvarez, A. Mallya, J. Kautz, and P. Molchanov, "A-vit: Adaptive tokens for efficient vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10809–10818, 2022.

[43] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman, "Token merging: Your vit but faster," in *The Eleventh International Conference on Learning Representations*, 2022.

[44] M. Fayyaz, S. A. Koohpayegani, F. R. Jafari, S. Sengupta, H. R. V. Joze, E. Sommerlade, H. Pirsiavash, and J. Gall, "Adaptive token sampling for efficient vision transformers," in *European Conference on Computer Vision*, pp. 396–414, Springer, 2022.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[46] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *ICML*, vol. 2, p. 4, 2021.

[47] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, A. Gupta, X. Gu, A. G. Hauptmann, *et al.*, "Language model beats diffusion–tokenizer is key to visual generation," *arXiv preprint arXiv:2310.05737*, 2023.

[48] N. Shazeer, "Fast transformer decoding: One write-head is all you need," 2019.

[49] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai, "Gqa: Training generalized multi-query transformer models from multi-head checkpoints," *arXiv preprint arXiv:2305.13245*, 2023.

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

[51] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," *arXiv preprint arXiv:1503.03585*, 2015.

[52] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[53] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[54] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, and J. Zhu, "All are worth words: A vit backbone for diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[55] S. Gao, P. Zhou, M.-M. Cheng, and S. Yan, "Masked diffusion transformer is a strong image synthesizer," *arXiv preprint arXiv:2303.14389*, 2023.

[56] A. Hatamizadeh, J. Song, G. Liu, J. Kautz, and A. Vahdat, "Diffit: Diffusion vision transformers for image generation," *arXiv preprint arXiv:2312.02139*, 2023.

[57] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[58] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," *arXiv preprint arXiv:2202.00512*, 2022.

[59] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 2249–2281, 2022.

[60] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[61] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[62] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *arXiv*, 2020.

[63] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022.

[64] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja, *et al.*, "Multitask prompted training enables zero-shot task generalization," *arXiv preprint arXiv:2110.08207*, 2021.

[65] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

[66] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, 2022.

[67] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*, pp. 4904–4916, PMLR, 2021.

[68] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *arXiv preprint arXiv:2205.01917*, 2022.

[69] S. Yan, T. Zhu, Z. Wang, Y. Cao, M. Zhang, S. Ghosh, Y. Wu, and J. Yu, "Video-text modeling with zero-shot transfer from contrastive captioners," *arXiv preprint arXiv:2212.04979*, 2022.

[70] H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang, *et al.*, "mplug-2: A modularized multi-modal foundation model across text, image and video," *arXiv preprint arXiv:2302.00402*, 2023.

[71] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "Git: A generative image-to-text transformer for vision and language," *arXiv preprint arXiv:2205.14100*, 2022.

[72] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Zero-shot video question answering via frozen bidirectional language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 124–141, 2022.

[73] Y. Li, "A practical survey on zero-shot prompt design for in-context learning," in *Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings*, RANLP, INCOMA Ltd., Shoumen, BULGARIA, 2023.

[74] B. Chen, Z. Zhang, N. Langrené, and S. Zhu, "Unleashing the potential of prompt engineering in large language models: a comprehensive review," *arXiv preprint arXiv:2310.14735*, 2023.

[75] S. Pitis, M. R. Zhang, A. Wang, and J. Ba, "Boosted prompt ensembles for large language models," 2023.

[76] Y. Hao, Z. Chi, L. Dong, and F. Wei, "Optimizing prompts for text-to-image generation," 2023.

[77] S. Huang, B. Gong, Y. Pan, J. Jiang, Y. Lv, Y. Li, and D. Wang, "Vop: Text-video co-operative prompt tuning for cross-modal retrieval," 2023.

[78] J. Z. Wu, Y. Ge, X. Wang, W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," 2023.

[79] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7086–7096, June 2022.

[80] X. Chen, Y. Wang, L. Zhang, S. Zhuang, X. Ma, J. Yu, Y. Wang, D. Lin, Y. Qiao, and Z. Liu, "Seine: Short-to-long video diffusion model for generative transition and prediction," 2023.

[81] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," 2024.

[82] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," 2018.

[83] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro, "Few-shot video-to-video synthesis," 2019.

[84] D. J. Zhang, D. Li, H. Le, M. Z. Shou, C. Xiong, and D. Sahoo, "Moonshot: Towards controllable video generation and editing with multimodal conditions," 2024.

[85] L. Zhuo, G. Wang, S. Li, W. Wu, and Z. Liu, "Fast-vid2vid: Spatial-temporal compression for video-to-video synthesis," 2022.

[86] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," 2021.

[87] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, 2021.

[88] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision*, pp. 709–727, Springer, 2022.

[89] OpenAI, "Introducing chatgpt," 2023.

[90] OpenAI, "Gpt-4v(ision) system card," 2023.

[91] Y. Huang and L. Sun, "Harnessing the power of chatgpt in fake news: An in-depth exploration in generation, detection and explanation," 2023.

[92] C. Chen and K. Shu, "Can llm-generated misinformation be detected?," 2023.

[93] Z. Liu, Y. Huang, X. Yu, L. Zhang, Z. Wu, C. Cao, H. Dai, L. Zhao, Y. Li, P. Shu, F. Zeng, L. Sun, W. Liu, D. Shen, Q. Li, T. Liu, D. Zhu, and X. Li, "Deid-gpt: Zero-shot medical text de-identification by gpt-4," 2023.

[94] J. Yao, X. Yi, X. Wang, Y. Gong, and X. Xie, "Value fulcra: Mapping large language models to the multidimensional spectrum of basic human values," 2023.

[95] Y. Huang, Q. Zhang, P. S. Y, and L. Sun, "Trustgpt: A benchmark for trustworthy and responsible large language models," 2023.

[96] L. Sun, Y. Huang, H. Wang, S. Wu, Q. Zhang, C. Gao, Y. Huang, W. Lyu, Y. Zhang, X. Li, Z. Liu, Y. Liu, Y. Wang, Z. Zhang, B. Kailkhura, C. Xiong, C. Xiao, C. Li, E. Xing, F. Huang, H. Liu, H. Ji, H. Wang, H. Zhang, H. Yao, M. Kellis, M. Zitnik, M. Jiang, M. Bansal, J. Zou, J. Pei, J. Liu, J. Gao, J. Han, J. Zhao, J. Tang, J. Wang, J. Mitchell, K. Shu, K. Xu, K.-W. Chang, L. He, L. Huang, M. Backes, N. Z. Gong, P. S. Yu, P.-Y. Chen, Q. Gu, R. Xu, R. Ying, S. Ji, S. Jana, T. Chen, T. Liu, T. Zhou, W. Wang, X. Li, X. Zhang, X. Wang, X. Xie, X. Chen, X. Wang, Y. Liu, Y. Ye, Y. Cao, Y. Chen, and Y. Zhao, "Trustllm: Trustworthiness in large language models," 2024.

[97] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, D. Forsyth, and D. Hendrycks, "Harmbench: A standardized evaluation framework for automated red teaming and robust refusal," 2024.

[98] Y. Wang, H. Li, X. Han, P. Nakov, and T. Baldwin, "Do-not-answer: A dataset for evaluating safeguards in llms," 2023.

[99] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, *et al.*, "Decodingtrust: A comprehensive assessment of trustworthiness in gpt models," *arXiv preprint arXiv:2306.11698*, 2023.

[100] Z. Zhang, L. Lei, L. Wu, R. Sun, Y. Huang, C. Long, X. Liu, X. Lei, J. Tang, and M. Huang, "Safetybench: Evaluating the safety of large language models with multiple choice questions," 2023.

[101] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang, ""' do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models," *arXiv preprint arXiv:2308.03825*, 2023.

[102] X. Liu, N. Xu, M. Chen, and C. Xiao, "Autodan: Generating stealthy jailbreak prompts on aligned large language models," *arXiv preprint arXiv:2310.04451*, 2023.

[103] S. Zhu, R. Zhang, B. An, G. Wu, J. Barrow, Z. Wang, F. Huang, A. Nenkova, and T. Sun, "Autodan: Interpretable gradient-based adversarial attacks on large language models," 2023.

[104] A. Zhou, B. Li, and H. Wang, "Robust prompt optimization for defending language models against jailbreaking attacks," *arXiv preprint arXiv:2401.17263*, 2024.

[105] X. Guo, F. Yu, H. Zhang, L. Qin, and B. Hu, "Cold-attack: Jailbreaking llms with stealthiness and controllability," 2024.

[106] A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: How does llm safety training fail?," *arXiv preprint arXiv:2307.02483*, 2023.

[107] Z. Niu, H. Ren, X. Gao, G. Hua, and R. Jin, "Jailbreaking attack against multimodal large language model," 2024.

[108] H. Liu, W. Xue, Y. Chen, D. Chen, X. Zhao, K. Wang, L. Hou, R. Li, and W. Peng, "A survey on hallucination in large vision-language models," 2024.

[109] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob, D. Manocha, and T. Zhou, "Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models," 2023.

[110] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," 2023.

[111] Y. Huang, J. Shi, Y. Li, C. Fan, S. Wu, Q. Zhang, Y. Liu, P. Zhou, Y. Wan, N. Z. Gong, *et al.*, "Metatool benchmark for large language models: Deciding whether to use tools and which to use," *arXiv preprint arXiv:2310.03128*, 2023.

[112] F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, and L. Wang, "Mitigating hallucination in large multi-modal models via robust instruction tuning," 2023.

[113] L. Wang, J. He, S. Li, N. Liu, and E.-P. Lim, "Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites," in *International Conference on Multimedia Modeling*, pp. 32–45, Springer, 2024.

[114] Y. Zhou, C. Cui, J. Yoon, L. Zhang, Z. Deng, C. Finn, M. Bansal, and H. Yao, "Analyzing and mitigating object hallucination in large vision-language models," *arXiv preprint arXiv:2310.00754*, 2023.

[115] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed, "Bias and fairness in large language models: A survey," *arXiv preprint arXiv:2309.00770*, 2023.

[116] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, and X. He, "Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation," *arXiv preprint arXiv:2305.07609*, 2023.

[117] Y. Liang, L. Cheng, A. Payani, and K. Shu, "Beyond detection: Unveiling fairness vulnerabilities in abusive language models," 2023.

[118] F. Friedrich, P. Schramowski, M. Brack, L. Struppek, D. Hintersdorf, S. Luccioni, and K. Kersting, "Fair diffusion: Instructing text-to-image generation models on fairness," *arXiv preprint arXiv:2302.10893*, 2023.

[119] R. Liu, C. Jia, J. Wei, G. Xu, L. Wang, and S. Vosoughi, "Mitigating political bias in language models through reinforced calibration," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 14857–14866, May 2021.

[120] R. K. Mahabadi, Y. Belinkov, and J. Henderson, "End-to-end bias mitigation by modelling biases in corpora," 2020.

[121] N. Mireshghallah, H. Kim, X. Zhou, Y. Tsvetkov, M. Sap, R. Shokri, and Y. Choi, "Can llms keep a secret? testing privacy implications of language models via contextual integrity theory," *arXiv preprint arXiv:2310.17884*, 2023.

[122] R. Plant, V. Giuffrida, and D. Gkatzia, "You are what you write: Preserving privacy in the era of large language models," *arXiv preprint arXiv:2204.09391*, 2022.

[123] H. Li, Y. Chen, J. Luo, Y. Kang, X. Zhang, Q. Hu, C. Chan, and Y. Song, "Privacy in large language models: Attacks, defenses and future directions," 2023.

[124] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the opportunities and risks of foundation models," 2022.

[125] T. Shen, R. Jin, Y. Huang, C. Liu, W. Dong, Z. Guo, X. Wu, Y. Liu, and D. Xiong, "Large language model alignment: A survey," *arXiv preprint arXiv:2309.15025*, 2023.

[126] X. Liu, X. Lei, S. Wang, Y. Huang, Z. Feng, B. Wen, J. Cheng, P. Ke, Y. Xu, W. L. Tam, X. Zhang, L. Sun, H. Wang, J. Zhang, M. Huang, Y. Dong, and J. Tang, "Alignbench: Benchmarking chinese alignment of large language models," 2023.

[127] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," 2023.

[128] T. Yu, Y. Yao, H. Zhang, T. He, Y. Han, G. Cui, J. Hu, Z. Liu, H.-T. Zheng, M. Sun, and T.-S. Chua, "Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback," 2023.

[129] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing.," *Neurocomputing*, p. 126232, 2023.

[130] OpenAI, "Sora safety." https://openai.com/sora#safety, 2024.

[131] Z. Fei, X. Shen, D. Zhu, F. Zhou, Z. Han, S. Zhang, K. Chen, Z. Shen, and J. Ge, "Lawbench: Benchmarking legal knowledge of large language models," *arXiv preprint arXiv:2309.16289*, 2023.

[132] Y. Li, Y. Huang, Y. Lin, S. Wu, Y. Wan, and L. Sun, "I think, therefore i am: Benchmarking awareness of large language models using awarebench," 2024.

[133] J. Zhu, H. Yang, H. He, W. Wang, Z. Tuo, W.-H. Cheng, L. Gao, J. Song, and J. Fu, "Moviefactory: Automatic movie creation from text using large generative models for language and images," *arXiv preprint arXiv:2306.07257*, 2023.

[134] J. Zhu, H. Yang, W. Wang, H. He, Z. Tuo, Y. Yu, W.-H. Cheng, L. Gao, J. Song, J. Fu, *et al.*, "Mobilevidfactory: Automatic diffusion-based social media video generation for mobile devices from text," in *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9371–9373, 2023.

[135] S. Zhuang, K. Li, X. Chen, Y. Wang, Z. Liu, Y. Qiao, and Y. Wang, "Vlogger: Make your dream a vlog," *arXiv preprint arXiv:2401.09414*, 2024.

[136] R. Feng, W. Weng, Y. Wang, Y. Yuan, J. Bao, C. Luo, Z. Chen, and B. Guo, "Ccedit: Creative and controllable video editing via diffusion models," *arXiv preprint arXiv:2309.16496*, 2023.

[137] J. Xing, M. Xia, Y. Liu, Y. Zhang, Y. Zhang, Y. He, H. Liu, H. Chen, X. Cun, X. Wang, *et al.*, "Make-your-video: Customized video generation using textual and structural guidance," *arXiv preprint arXiv:2306.00943*, 2023.

[138] Y. Guo, C. Yang, A. Rao, Y. Wang, Y. Qiao, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023.

[139] Y. He, M. Xia, H. Chen, X. Cun, Y. Gong, J. Xing, Y. Zhang, X. Wang, C. Weng, Y. Shan, *et al.*, "Animate-a-story: Storytelling with retrieval-augmented video generation," *arXiv preprint arXiv:2307.06940*, 2023.

[140] H. Ni, C. Shi, K. Li, S. X. Huang, and M. R. Min, "Conditional image-to-video generation with latent flow diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18444–18455, 2023.

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

[142] Y. Hu, C. Luo, and Z. Chen, "Make it move: controllable image-to-video generation with text descriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18219–18228, 2022.

[143] K. Mei and V. Patel, "Vidm: Video implicit diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 9117–9125, 2023.

[144] S. Yu, K. Sohn, S. Kim, and J. Shin, "Video probabilistic diffusion models in projected latent space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18456–18466, 2023.

[145] K. Su, K. Qian, E. Shlizerman, A. Torralba, and C. Gan, "Physics-driven diffusion models for impact sound synthesis from videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9749–9759, 2023.

[146] S. Li, W. Dong, Y. Zhang, F. Tang, C. Ma, O. Deussen, T.-Y. Lee, and C. Xu, "Dance-to-music generation with encoder-based textual inversion of diffusion models," *arXiv preprint arXiv:2401.17800*, 2024.

[147] A. Awasthi, J. Nizam, S. Zare, S. Ahmad, M. J. Montalvo, N. Varadarajan, B. Roysam, and H. V. Nguyen, "Video diffusion models for the apoptosis forcasting," *bioRxiv*, pp. 2023–11, 2023.

[148] A. Bozorgpour, Y. Sadegheih, A. Kazerouni, R. Azad, and D. Merhof, "Dermosegdiff: A boundary-aware segmentation diffusion model for skin lesion delineation," in *International Workshop on PRedictive Intelligence In MEdicine*, pp. 146–158, Springer, 2023.

[149] A. Flaborea, L. Collorone, G. M. D. di Melendugno, S. D'Arrigo, B. Prenkaj, and F. Galasso, "Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10318–10329, 2023.

[150] J. Wu, R. Fu, H. Fang, Y. Zhang, and Y. Xu, "Medsegdiff-v2: Diffusion based medical image segmentation with transformer," *arXiv preprint arXiv:2301.11798*, 2023.

[151] G. J. Chowdary and Z. Yin, "Diffusion transformer u-net for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 622–631, Springer, 2023.

[152] I. Kapelyukh, V. Vosylius, and E. Johns, "Dall-e-bot: Introducing web-scale diffusion models to robotics," *IEEE Robotics and Automation Letters*, 2023.

[153] W. Liu, T. Hermans, S. Chernova, and C. Paxton, "Structdiffusion: Object-centric diffusion for semantic rearrangement of novel objects," in *Workshop on Language and Robotics at CoRL 2022*, 2022.

[154] M. Janner, Y. Du, J. B. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," *arXiv preprint arXiv:2205.09991*, 2022.

[155] A. Ajay, Y. Du, A. Gupta, J. Tenenbaum, T. Jaakkola, and P. Agrawal, "Is conditional generative modeling all you need for decision-making?," *arXiv preprint arXiv:2211.15657*, 2022.

[156] J. Carvalho, A. T. Le, M. Baierl, D. Koert, and J. Peters, "Motion planning diffusion: Learning and planning of robot motions with diffusion models," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1916–1923, IEEE, 2023.

[157] X. Gu, C. Wen, J. Song, and Y. Gao, "Seer: Language instructed video prediction with latent diffusion models," *arXiv preprint arXiv:2303.14897*, 2023.

[158] Z. Chen, S. Kiami, A. Gupta, and V. Kumar, "Genaug: Retargeting behaviors to unseen situations via generative augmentation," *arXiv preprint arXiv:2302.06671*, 2023.

[159] Z. Mandi, H. Bharadhwaj, V. Moens, S. Song, A. Rajeswaran, and V. Kumar, "Cacti: A framework for scalable multi-task multi-scene visual imitation learning," *arXiv preprint arXiv:2212.05711*, 2022.

[160] T. Chen, L. Li, S. Saxena, G. Hinton, and D. J. Fleet, "A generalist framework for panoptic segmentation of images and videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 909–919, 2023.

[161] W. Harvey, S. Naderiparizi, V. Masrani, C. Weilbach, and F. Wood, "Flexible diffusion modeling of long videos," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27953–27965, 2022.

[162] A. Gupta, S. Tian, Y. Zhang, J. Wu, R. Martín-Martín, and L. Fei-Fei, "Maskvit: Masked visual pre-training for video prediction," *arXiv preprint arXiv:2206.11894*, 2022.

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

[164] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

[165] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu, and J. Feng, "Magicvideo: Efficient video generation with latent diffusion models," *arXiv preprint arXiv:2211.11018*, 2022.

[166] S. Ge, T. Hayes, H. Yang, X. Yin, G. Pang, D. Jacobs, J.-B. Huang, and D. Parikh, "Long video generation with time-agnostic vqgan and time-sensitive transformer," in *European Conference on Computer Vision*, pp. 102–118, Springer, 2022.

[167] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, and D. Erhan, "Phenaki: Variable length video generation from open domain textual description," *arXiv preprint arXiv:2210.02399*, 2022.

[168] 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*, pp. 7346–7356, 2023.

[169] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," *arXiv preprint arXiv:2303.13439*, 2023.

[170] Z. Luo, D. Chen, Y. Zhang, Y. Huang, L. Wang, Y. Shen, D. Zhao, J. Zhou, and T. Tan, "Videofusion: Decomposed diffusion models for high-quality video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10209–10218, 2023.

[171] A. Jabri, D. Fleet, and T. Chen, "Scalable adaptive computation for iterative generation," *arXiv preprint arXiv:2212.11972*, 2022.

[172] L. Lian, B. Shi, A. Yala, T. Darrell, and B. Li, "Llm-grounded video diffusion models," *arXiv preprint arXiv:2309.17444*, 2023.

[173] E. Molad, E. Horwitz, D. Valevski, A. R. Acha, Y. Matias, Y. Pritch, Y. Leviathan, and Y. Hoshen, "Dreamix: Video diffusion models are general video editors," *arXiv preprint arXiv:2302.01329*, 2023.

[174] J. H. Liew, H. Yan, J. Zhang, Z. Xu, and J. Feng, "Magicedit: High-fidelity and temporally coherent video editing," *arXiv preprint arXiv:2308.14749*, 2023.

[175] W. Chen, J. Wu, P. Xie, H. Wu, J. Li, X. Xia, X. Xiao, and L. Lin, "Control-a-video: Controllable text-to-video generation with diffusion models," *arXiv preprint arXiv:2305.13840*, 2023.

[176] W. Chai, X. Guo, G. Wang, and Y. Lu, "Stablevideo: Text-driven consistency-aware diffusion video editing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23040–23050, 2023.

[177] S. Yang, Y. Zhou, Z. Liu, and C. C. Loy, "Rerender a video: Zero-shot text-guided video-to-video translation," *arXiv preprint arXiv:2306.07954*, 2023.

[178] D. Ceylan, C.-H. P. Huang, and N. J. Mitra, "Pix2video: Video editing using image diffusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23206–23217, 2023.

[179] B. Qin, J. Li, S. Tang, T.-S. Chua, and Y. Zhuang, "Instructvid2vid: Controllable video editing with natural language instructions," *arXiv preprint arXiv:2305.12328*, 2023.

[180] D. Liu, Q. Li, A.-D. Dinh, T. Jiang, M. Shah, and C. Xu, "Diffusion action segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10139–10149, 2023.

[181] R. Feng, Y. Gao, T. H. E. Tse, X. Ma, and H. J. Chang, "Diffpose: Spatiotemporal diffusion model for video-based human pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14861–14872, 2023.

[182] L. Yu, Y. Cheng, K. Sohn, J. Lezama, H. Zhang, H. Chang, A. G. Hauptmann, M.-H. Yang, Y. Hao, I. Essa, *et al.*, "Magvit: Masked generative video transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023.

[183] Z. Li, R. Tucker, N. Snavely, and A. Holynski, "Generative image dynamics," *arXiv preprint arXiv:2309.07906*, 2023.

[184] EasyWithAI, "Zeroscope - ai text-to-video model." https://easywithai.com/ai-video-generators/zeroscope/, 2023.

[185] R. Girdhar, M. Singh, A. Brown, Q. Duval, S. Azadi, S. S. Rambhatla, A. Shah, X. Yin, D. Parikh, and I. Misra, "Emu video: Factorizing text-to-video generation by explicit image conditioning," *arXiv preprint arXiv:2311.10709*, 2023.

[186] Y. Zeng, G. Wei, J. Zheng, J. Zou, Y. Wei, Y. Zhang, and H. Li, "Make pixels dance: High-dynamic video generation," *arXiv preprint arXiv:2311.10982*, 2023.

[187] A. Gupta, L. Yu, K. Sohn, X. Gu, M. Hahn, L. Fei-Fei, I. Essa, L. Jiang, and J. Lezama, "Photorealistic video generation with diffusion models," *arXiv preprint arXiv:2312.06662*, 2023.

[188] B. Wu, C.-Y. Chuang, X. Wang, Y. Jia, K. Krishnakumar, T. Xiao, F. Liang, L. Yu, and P. Vajda, "Fairy: Fast parallelized instruction-guided video-to-video synthesis," *arXiv preprint arXiv:2312.13834*, 2023.

[189] D. Kondratyuk, L. Yu, X. Gu, J. Lezama, J. Huang, R. Hornung, H. Adam, H. Akbari, Y. Alon, V. Birodkar, *et al.*, "Videopoet: A large language model for zero-shot video generation," *arXiv preprint arXiv:2312.14125*, 2023.

[190] J. Wu, X. Li, C. Si, S. Zhou, J. Yang, J. Zhang, Y. Li, K. Chen, Y. Tong, Z. Liu, *et al.*, "Towards language-driven video inpainting via multimodal large language models," *arXiv preprint arXiv:2401.10226*, 2024.

[191] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada, A. Ephrat, J. Hur, Y. Li, T. Michaeli, *et al.*, "Lumiere: A space-time diffusion model for video generation," *arXiv preprint arXiv:2401.12945*, 2024.

A Related Works

We show some related works about the video generation tasks in Table 1.

Table 1: Summary of Video Generation.

Model name	Year	Backbone	Task	Group
Imagen Video[29]	2022	Diffusion	Generation	Google
Pix2Seq-D[160]	2022	Diffusion	Segmentation	Google Deepmind
FDM[161]	2022	Diffusion	Prediction	UBC
MaskViT[162]	2022	Masked Vision Models	Prediction	Stanford, Salesforce
CogVideo[163]	2022	Auto-regressive	Generation	THU
Make-a-video[164]	2022	Diffusion	Generation	Meta
MagicVideo[165]	2022	Diffusion	Generation	ByteDance
TATS[166]	2022	Auto-regressive	Generation	University of Maryland, Meta
Phenaki[167]	2022	Masked Vision Models	Generation	Google Brain
Gen-1[168]	2023	Diffusion	Generation, Editing	RunwayML
LFDM[140]	2023	Diffusion	Generation	PSU, UCSD
Text2video-Zero[169]	2023	Diffusion	Generation	Picsart
Video Fusion[170]	2023	Diffusion	Generation	USAC, Alibaba
PYoCo[34]	2023	Diffusion	Generation	Nvidia
Video LDM[36]	2023	Diffusion	Generation	University of Maryland, Nvidia
RIN[171]	2023	Diffusion	Generation	Google Brain
LVD[172]	2023	Diffusion	Generation	UCB
Dreamix[173]	2023	Diffusion	Editing	Google
MagicEdit[174]	2023	Diffusion	Editing	ByteDance
Control-A-Video[175]	2023	Diffusion	Editing	Sun Yat-Sen University
StableVideo[176]	2023	Diffusion	Editing	ZJU, MSRA
Tune-A-Video[78]	2023	Diffusion	Editing	NUS
Rerender-A-Video[177]	2023	Diffusion	Editing	NTU
Pix2Video[178]	2023	Diffusion	Editing	Adobe, UCL
InstructVid2Vid[179]	2023	Diffusion	Editing	ZJU
DiffAct[180]	2023	Diffusion	Action Detection	University of Sydney
DiffPose[181]	2023	Diffusion	Pose Estimation	Jilin University
MAGVIT[182]	2023	Masked Vision Models	Generation	Google
AnimateDiff[138]	2023	Diffusion	Generation	CUHK
MAGVIT V2[47]	2023	Masked Vision Models	Generation	Google
Generative Dynamics[183]	2023	Diffusion	Generation	Google
VideoCrafter[81]	2023	Diffusion	Generation	Tencent
Zeroscope[184]	2023	-	Generation	EasyWithAI
ModelScope	2023	-	Generation	Damo
Gen-2[23]	2023	-	Generation	RunwayML
Pika[22]	2023	-	Generation	Pika Labs
Emu Video[185]	2023	Diffusion	Generation	Meta
PixelDance[186]	2023	Diffusion	Generation	ByteDance
Stable Video Diffusion[27]	2023	Diffusion	Generation	Stability AI
W.A.L.T[187]	2023	Diffusion	Generation	Stanford, Google
Fairy[188]	2023	Diffusion	Generation, Editing	Meta
VideoPoet[189]	2023	Auto-regressive	Generation, Editing	Google
LGVI[190]	2024	Diffusion	Editing	PKU, NTU
Lumiere[191]	2024	Diffusion	Generation	Google
Sora[3]	2024	Diffusion	Generation, Editing	OpenAI