The Diffusion Revolution: Sora, Open-Sora, and the Future of Generative Video

Brenden Noblitt*
*Department of Computer Science and Information Systems, East Texas A&M University, Commerce, TX, USA
Email: bnoblitt@myleo.tamuc.edu

***Abstract*— This paper surveys the evolution of generative models in computer vision, with a specific focus on the shift from adversarial approaches to diffusion-based techniques for image and video synthesis. Beginning with foundational models like GANs and VQ-VAE, the work traces milestones in the emergence of diffusion models, such as DDPM, Latent Diffusion, and Video Diffusion models, leading up to modern systems like OpenAI's Sora and the open-source project Open-Sora. This paper also explores real-world applications of video generation technology in domains such as optometry, interior design, and mobile devices. This paper concludes with potential research directions for the future, such as controllable video generation, token compression, and incorporated human feedback. The review aims to contextualize Sora and Open-Sora's technical innovations while charting a path for more accessible, controllable, and efficient generative video systems.**

Keywords—Video Generation, Image Generation, Generative AI, Diffusion Models, Sora, Open-Sora, Multi-Modal Models, Generative Media

I. Introduction

Generative AI has become deeply intertwined with the field of computer vision, transforming how images and videos are synthesized, manipulated, and understood. What began as early attempts to generate low-resolution images using CNNs and adversarial learning has now evolved into systems capable of producing photorealistic, temporally coherent video sequences from natural language prompts.

This transformation has been driven by key breakthroughs in multiple areas, such as architectural innovation, training scalability, and a fundamental shift from adversarial to diffusion-based models. Models like DALL-E, Imagen, and Stable Diffusion reshaped image synthesis by introducing controlled, high-resolution generation in compressed latent spaces. In parallel, text-to-video models, such as OpenAI's Sora, have pushed the frontier of generative AI from still frames into dynamic, high-fidelity motion with temporal and physical realism.

This paper follows the evolution of these generative systems, focusing in particular on the rise of diffusion models and their impact on image and video generation. We review the historical progression from early diffusion methods to high-performance models like Sora, and analyze the emergence of open-source alternatives such as Open-Sora Plan, Open-Sora 2.0, and Open-Sora (STDiT). These models not only democratize access to advanced video synthesis capabilities but also reflect critical research themes, such as modularity, efficiency, temporal coherence, and controllability.

In doing so, the aim is to highlight both the technical milestones and the broader research trajectory that define today's generative video landscape. This paper will also explore real-world applications, limitations of current systems, and promising directions for future research.

II. Background Information

While modern diffusion-based generation models like Sora capture widespread attention for their stunning outputs, they are built upon years of foundational research in generative modeling and computer vision. Understanding these systems requires the context of the broader evolution of generative approaches in computer vision. From the early development of variational autoencoders (VAEs) and generative adversarial networks (GANs) to the more recent shift towards diffusion models and latent-space training strategies, the landscape has been shaped by continual advances in both model architecture and generative fidelity.

In this section, we provide an overview of the core developments that made modern models like Sora and Open-Sora possible. We begin by examining early diffusion models and their transition from pixel-space to latent-space generation. We then explore how these techniques extend to video through architectural innovations like 3D U-Nets and spatial-temporal transformers. This timeline sets the stage for understanding the technical motivations behind Open-Sora's design choices and how each generation of models has built upon the limitations and capabilities of its predecessors.

A. Early Image Generation (2014-2016)

Generative Adversarial Nets (GANs) are the first technology to discuss and a major component of video generation technology. These innovations helped to create the foundation for image and video generation and ultimately led to various discoveries in computer vision and generative AI. GANs are structured around the idea of a new framework for generative modeling. This framework attempts to use new strategies for optimizing computations, such as maximum likelihood estimations and sampling-based methods. Previous methods for generating high-quality images and videos were expensive due to computational constraints. The GAN paradigm involves a two-player game between a generative model and an "adversary", a discriminative model that is taught to determine if the selected sample is from the model or data distribution. This method enables both the generative and discriminative models to compete against each other and continually improve until their outputs are undistinguishable from each other. Within this paradigm, a minmax optimization function is also used to improve training. It is also demonstrated that GANs have potential within unsupervised learning. The adversarial framework was tested against popular datasets, such as MNIST and the Toronto Face Database, and produced results competitive with existing models.

Some limitations of GANs include training instability, where vanishing gradients can pose issues if the adversarial model becomes too strong. The model itself may also produce limited variety, which can cause issues when attempting to capture the full diversity of the data distribution. Finally, GANs do not provide a tractable way to compute likelihoods, which makes evaluation difficult. GANs paved the way for the subfield of generative modeling. Research in GANs set the foundation for diffusion models and video synthesis, adversarial training, text-to-image synthesis, and the overall shift to adversarial paradigms. Overall, GANs ultimately led to the techniques used in modern-day text-to-image/video generation models, such as SORA [1].

Next, we will discuss image-to-image translation with conditional adversarial networks, made popular by the Pix2Pix model. This approach to synthesizing images involved solving some problems with image-to-image translation, such as adding color to a grayscale image or turning a sketch into a full photo. These translation tasks required substantial manual effort by the engineers and researchers using these methods, as tasks like labeling and optimizing loss functions were required. This altered approach suggests using a conditional GAN framework, conditioning both the generator and discriminator on the image and enabling supervised image translation. The GAN will learn the loss that tries to classify whether the output image is real or fake, while the generative model is trained to minimize the loss. This approach also alters the optimized minmax function, combining cGAN loss with L1 loss to encourage both realism and correctness in the model output. A specific U-Net generator and PatchGAN discriminator are used to construct the GAN. The U-Net architecture improves output fidelity by allowing low-level features to bypass deeper levels of the network, while the PatchGAN discriminator classifies in patches instead of the entire image at once, which improves quality in the details of the image.

Some limitations for this approach include performance dependencies on paired training data, limited generalization with task-specific knowledge, and inconsistencies with visual artifacts in the output. While these limitations can pose some issues, this approach shows how improvement in GAN technology can be applied in image-to-image translation. The approach behind Pix2Pix demonstrates how GANs can be shaped by their conditioning inputs. This idea is used in modern text-to-image/video generation models like SORA, which conditions on prompts. This approach also sets some of the foundations for future diffusion and video-to-video synthesis models with its paired translation functionality [2].

B. Autoencoders and Discrete Latents (2019-2020)

While GANs and the Pix2Pix model provided revolutionary innovations to the field, these research ambitions carried over into the usage of autoencoders and discrete latents in image and video generation. Autoencoders introduced the concept of models learning efficient data representation in an unsupervised manner, while discrete latents allowed image data to be represented as a small number of discrete tokens.

To start, we will discuss neural discrete representation learning. Learning a useful representation without supervision in the field of machine learning remains a challenge. This approach attempts to train a model on discrete latent representations, which assist in working with symbolic reasoning and efficient generative modeling. This approach also attempts to improve quality and efficiency in latent variable models. Continuous latent spaces do not always align well with true data distribution, which is used in variational autoencoders. This approach introduces the Vector Quantized Variational Autoencoder (VQ-VAE), which is a novel method combining neural networks with a learned discrete codebook. The encoder maps input to the nearest entry in a learned codebook while the decoder reconstructs the data from the quantized latent vector. VQ-VAEs differ from VAEs in two ways: the encoder network outputs discrete, rather than continuous, codes, and the prior is learnt rather than static. Some other contributions include the introduction of a straight-through estimator, which assists with quantization, and the demonstration that discrete latent spaces enable high-quality image reconstruction.

Some limitations with this approach include unused codebook entries, which can create wasted resources and misleading interpretations. Quantization noise can also impact the model, where the nearest codebook vector loses information, and fine details become lost. This approach also does not use explicit likelihood training, which can make evaluation more difficult. This approach pioneered discrete latent variable modeling. This directly influenced image/video synthesis methods, such as SORA and DALL-E, the usage of transformers over discrete latents, which enabled more compact, tokenized representations of long videos, and provided a bridge between vision and language [3].

Next, we will discuss the Vector Quantized Variational Autoencoder (VQ-VAE). While VQ-VAE created innovations in image generation technology, an effort was made to improve on this foundational model through VQ-VAE-2. The intent was to improve coherence and fidelity in the outputs compared to the previous model, as well as create more diversity in the samples. In turn, the overall goal was to build a model that could produce diverse and high-quality images using discrete latent variables while keeping training stable and efficient. This approach introduces VQ-VAE-2, which uses a multi-scale latent hierarchy where lower-resolution latents model global structure and higher-resolution latents refine local details. This approach also uses PixelCNN and PixelSNAIL over the discrete latent variables at each level, which creates strong multi-scale generational modeling. These modifications introduce high-quality image generation compared to VQ-VAE-1 and achieve results on par with state-of-the-art GANs without adversarial training.

The training for VQ-VAE-2 requires multiple stages of training, which can utilize more computing resources. Another limitation is in the autoregressive sampling, which can be slow at high resolutions. Next, unlike some text-to-image models, this model does not condition semantic inputs. The last limitation to note is that the PixelCNN component may not scale efficiently to longer video sequences without optimization. VQ-VAE-2 helped set the stage for modern image and video generation technologies. Modern technologies like SORA and DALL-E use the hierarchical latent compression proposed with this model, which enables high detail without additional compute. This model also introduced the concept that multi-scale latent spaces can split into a global structure and local texture, which assists in video generation tasks for long videos [4].

C. Early Diffusion Models (2021-2022)

While autoencoders and discrete latents opened doors for more advanced image and video generation techniques, these technologies still had their limitations. GANs, using their adversarial approach, struggle with mode collapse and vanishing gradients. The next models we will discuss set the stage for diffusion model usage in image/video generation and propel generative technologies into modern perception.

First, we will discuss Denoising Diffusion Probabilistic Models (DDPM), a popular method that innovated the usage of diffusion models in image and video generation. This approach, using diffusion probabilistic models, attempts to close the gap between sample fidelity, a concept championed by GANs, and training stability and evaluation, a strength of likelihood-based models like VAEs and autoregressive models. This approach reconstructed diffusion models as score-based generative models. This involves using a denoising process as a learning step with the model and a forward process that gradually adds Gaussian noise over the timesteps, and a reverse process to remove it. This approach also uses a simple objective function, which uses a variational lower bound to optimize the model, but simplifies it to mean-squared errors between predicted and true noise. Ultimately, this approach demonstrated that models with similar evaluation metrics to state-of-the-art models can be trained without using adversarial training. It also showed the modularity of the diffusion process, which creates an opportunity for improvement within the general system.

Some limitations to this approach include slow sampling, as samples require many denoising steps, which can be slower than GANs. Training diffusion models is computationally intensive, especially at high resolutions. Finally, this approach lacks conditioning mechanisms, which makes it difficult to guide the model to video generation with text or other media. This approach revived interest in diffusion models, which are a foundational piece of modern AI technologies. Future text-to-image models, such as Imagen and Stable Diffusion, build directly on this approach by adding conditioning mechanisms. Modern video diffusion models, like SORA, use this architecture and adapt it to spatiotemporal data. Overall, DDPM created a shift in the research community from adversarial training methods to likelihood-based generative modeling methods [5].

Next, we will discuss high-resolution image synthesis with latent diffusion models, a method that was introduced using diffusion in latent space. The existing diffusion models at the time, such as DDPM, were powerful but computationally expensive, especially at high resolutions. Training powerful diffusion models could take five days just to generate 50k samples. This creates two consequences: First, obtaining massive computational resources is a task only attainable by a small fraction of the field. Second, evaluating an already trained model is expensive in time and memory since the same model architecture must run sequentially for many steps. In order to create a more efficient, scalable diffusion-based pipeline, a new model called Latent Diffusion Models (LDM) is created. This approach involves training diffusion models in a learned latent space, which exhibits better scaling properties concerning spatial dimensionality. This approach allows the universal autoencoding stage to take place once and be reused in multiple trainings. This approach also utilizes a modular approach, with the encoder/decoder mapping images to the latent space, the U-Net Diffusion Model operating on the latent representation, and the conditioning module enabling text/image conditioning, all working together. This approach also enables synthesis from text prompts without paired training data. Overall, the new model scores well on benchmarks like COCO and ImageNet, capable of generating 1024x1024 images with fine detail.

Some limitations with this approach include a dependency on encoder quality, where the fidelity of the generated images are bound by the expressiveness of the VQ-VAE latent space. Diffusion in latent space could also create some complications, such as image artifacts or loss of subtle detail. The modular approach could also create some issues with implementation and deployment. This approach influenced newer diffusion-based approaches, such as Stable Diffusion, which utilizes the same core architecture and is based on this approach. This approach also introduced ideas like cross-attention condition and diffusion in latent space, which are concepts used by text-to-image/video models, such as DALL-E and Imagen. Finally, the idea of generative models in latent space is still influencing video diffusion models today, where computational savings are valuable [6].

D. Video Diffusion Models (2022-2023)

These discoveries in diffusion-based technologies created many different research opportunities to harness this technology within video generation models in order to create quality video frames at scale. Next, we will discuss popular video generation discoveries that propelled the next generation of video generation models.

First, we will discuss a foundational concept called video diffusion models. This approach attempts to show a new model capable of generating realistic and temporally coherent videos as well as overcoming some of the limitations of previous video generation approaches, such as GANs. In order to attain these tasks, a diffusion model, originally used with images, is adapted to handle video. This approach demonstrates the first strong diffusion-based model for video generation at the time. It introduced a spatiotemporal U-Net architecture that can denoise entire video clips. There are two variants of the model, a Video Diffusion Model (VDM) that learns to generate full-resolution RGB video using a standard diffusion process across all frames. A Video Autoregressive Diffusion Model is also introduced, combining diffusion and autoregressive conditions to generate one frame at a time. This approach produces video with good motion coherence and achieves high scores on datasets such as UCF-101.

This approach utilizes expensive sampling, typically requiring hundreds of steps per video. Videos generated by this model were also relatively short, introducing some potential problems when using this in a production setting. This model also struggles with long-term dependencies as diffusion models typically on see a fixed-length window. This paper provided a foundation for modern models, such as VideoCrafter and SORA, all of which extend or optimize the VDM framework. This approach also demonstrated that diffusion models are not just for images but can also be used to model motion and action sequences, which are a crucial component of generative synthesis in computer vision. Ultimately, this approach helped to bridge the gap between text-to-image models and modern video generation models [7].

Next, we will discuss the Google Imagen model, a powerful model that pushed text-to-image technology to new applications. While this is an image-based model, the techniques used with generative AI influenced future video generation models in their usage of generative AI. While video diffusion techniques were being explored, text-to-image was another area being researched with the recent explosion of generative AI during this time. Imagen, a text-to-image diffusion model by Google, was one attempt at this research. Imagen focused on generating photorealistic, high-fidelity images from natural language prompts. It was also attempting to improve among existing models, such as OpenAI's DALL-E 2, which were limited in either image fidelity or text understanding. Google introduces Imagen to solve some of these issues. Imagen combines a frozen T5-XXL large language model for text encoding with a cascaded diffusion model for image generation. Diffusion models are used at multiple resolutions with super-resolution diffusion stages conditioned on text and previous outputs. Using the frozen T5 encoder also allows for high-quality embeddings without the need for joint training, which improves performance and scalability. Overall, Imagen's images are considered highly photorealistic by human evaluators and obtain FID scores superior to other models, such as DALL-E 2.

Some limitations of Imagen include its closed-source nature, where model weights and training data were not released to the public. Imagen also requires large-scale proprietary datasets and a large amount of compute, which may make reproducibility difficult. The frozen T5 encoder, while helpful, may limit adaptability and flexibility as language usage changes over time. Imagen is one of the initial efforts to combine generative AI with diffusion models to create state-of-the-art images from text-to-image methods. The diffusion pipelines offer a path towards scaling to high-resolution, high-fidelity outputs, methods that are used by modern models like SORA. Ultimately, Imagen demonstrated that diffusion models can be used in place of GANs for text-conditioned generation and paved the way for innovations in combining generative AI and computer vision [8].

E. Modern Video Generation Models (2023-Current)

With the improvements made on diffusion models used in video generation, as well as the innovations made to combine generative AI and video generation, new models were created that are still used today. These video generation models are recognized for their generated video quality, text-to-video capabilities, and their usage of generative AI.

First, we will discuss Meta's Make-A-Video, a foundational video generation model that eliminated the need for paired text-video training data. Technology in video generation was continuing to evolve; however, the challenge of finding large-scale text-video paired datasets was still present. Make-A-Video by Meta aimed to solve this issue by generating short, coherent videos from text prompts despite the absence of fully supervised text-video datasets. Make-A-Video also focused on leveraging existing image-generation models and datasets to bootstrap video generation capabilities. Make-A-Video is trained without any paired text-video data. It builds upon pre-trained text-to-image diffusion models and transfers the knowledge to the video domain. Some advantages of Make-A-Video include accelerated training time in the T2V model, no requirements for paired text-video data, and generated videos inherit the vastness of today's image generation models. The architecture of the model extends image models with spatiotemporal attention, which enables video synthesis from static image priors. Its training is also a multi-stage process that involves using image-text pairs for initial training and then training on unlabeled video data using self-supervised objectives.

Some limitations include the short duration of videos created by Make-A-Video, which are typically 16-32 frames. Videos can be low resolution and contain semantic drift at times. There is no fine-grained control in areas like object behavior or camera motion. Make-A-Video was also not released to the public due to safety concerns. Make-A-Video pioneered the strategy of bootstrapping from text-to-image models for video synthesis, a technique used by modern models such as SORA. Make-A-Video also introduced the concept that temporal consistency can be layered on top of strong image priors rather than requiring fully supervised video data. This model also helped lead the shift from adversarial video generation to diffusion-based techniques [9].

Next, we will discuss Tencent's VideoCrafter model, one of the first video generation models to introduce text-to-video capabilities. While Make-A-Video was an innovative model, it still struggled with challenges around resolution, scalability, and generalization. VideoCrafter by Tencent aimed to solve some of these problems. VideoCrafter is a general-purpose, open-source video generation framework based on diffusion models. It enables high-quality video generation through text-to-video synthesis, unconditional video generation, and image-to-video extension. It attempts to solve some of the issues with other video generation models and provide a modular, extensible pipeline for the community to build on. VideoCrafter also attempts to bring an open-source model to the community that researchers and developers can use and experiment with. VideoCrafter provides the VideoCrafter framework, based on latent diffusion for improved efficiency and scalability, which supports multiple video generation tasks. VideoCrafter is a Latent Video Diffusion Model (LVDM) which consists of a video VAE, a video latent diffusion model, and a 3D U-Net architecture. This allows for the components to operate on spatiotemporal latent representations extracted by the VAE. This supports unconditional and text-conditioned video generation. VideoCrafter uses a two-stage text-to-video pipeline. It uses a frozen text encoder, such as T5, to provide language conditioning. The diffusion model then generates video latents from text, which are then decoded into RGB videos by the decoder. Weights, training code, and evaluation scripts are also open-sourced for the community to use.

Some limitations of VideoCrafter include its short video lengths, typically spanning 16-32 frames. Initial releases of VideoCrafter had low resolutions but were scalable. The temporal consistency used is also limited, preventing long-range conditioning. Finally, initial evaluation metrics were not optimal; however, they are improved upon in VideoCrafter2. VideoCrafter emphasizes reproducibility through its open-source nature compared to closed models like Imagen and Make-A-Video. VideoCrafter also utilizes latent diffusion, which paves the way for scaling to higher resolutions and longer videos. It also demonstrates how its

unique architecture can create generalization across video generation tasks [10].

Next, we will discuss Runway's Gen-2 model, a video generation model that helped introduce the usage of multi-modal conditioning. Runway's Gen-2 vision model attempts to create more controllable and coherent video synthesis. It does this by enabling structure-guided video generation, supporting content conditioning, and producing temporarily photorealistic videos without requiring fully paired supervision. Existing models lacked structure awareness or required costly paired data, whereas Gen-2 attempts to decouple structure and content to allow flexibility. This approach provides a number of contributions. First, a dual-branch architecture is introduced, which separates the structure of a video from its content. A structure-guided U-Net backbone is introduced, where video is synthesized in the latent space using a spatiotemporal U-Net containing structure and content encoded and fused into the denoising process. Gen-2 also supports various input types, including text-to-video, image-to-video, and depth-to-video. Gen-2 also demonstrates the usage of image-text datasets for appearance learning and unlabeled video datasets for motion learning.

Some limitations of Gen-2 include short video duration, where videos are roughly 32 frames. Gen-2 was also limited to 512x512 resolution. Since Gen-2 is a diffusion-based model, computational costs remain high. Finally, the complexity created by merging multiple conditioning types, such as text and image, introduces challenges in weighting. Gen-2 demonstrated that a modular and controllable video generation pipeline can allow for the synthesis of long, coherent, and semantically aligned videos. Gen-2 also demonstrates the functionality of multi-modal conditioning, where structure governs motion, a concept used by modern models like SORA. Overall, Gen-2 displays a combination of controllability and coherence that allows more performant models to be built in the future [11].

III. Sora and Open-Sora

With the foundation of image and video diffusion models established, the release of Sora by OpenAI marked a significant milestone in the field. As one of the first systems to generate high-resolution, long-duration videos with physics-consistent motion and text-based control, Sora represented both a technical and conceptual leap forward. However, due to its closed-source nature, much about its internal architecture and training remains opaque. In response, a wave of open-source efforts emerged, most notably the Open-Sora project, which aims to replicate, democratize, and extend Sora-like capabilities through transparent, modular design. This section explores the capabilities of Sora, the motivations behind its open-source counterparts, and the architectural innovations that make models like Open-Sora Plan, Open-Sora 2.0, and Open-Sora (STDiT) technically and philosophically significant in the evolution of generative video analysis and generative AI.

A. Sora by OpenAI (2024)

OpenAI's Sora model attempts to revolutionize video generation by solving a set of problems. Generating realistic, temporally consistent videos from text prompts has been a challenge in the field. Building models that simulate aspects of the real world without a physics engine was another challenge. Finally, Sora focused on closing the gap between generative video quality and real-world generalization. Sora contributes a number of items to the field and industry. First, Sora reframes video diffusion models as implicit simulators of the physical world. This approach shows that video generation emerges from learning to predict outcomes. Sora can also generate photorealistic, high-resolution videos up to sixty seconds long. It can also handle long-range temporal dependencies, camera motion, and realistic physics, and can accept input in multiple modalities, such as text or images. Sora's multiple modalities also establish grounding in physics without needing explicitly labeled data. Finally, Sora's robust generalization allows for creative or fantastical scenarios to be formed.

Some limitations with Sora include its lack of public release. Lack of interactivity is also a limitation, as Sora generates from prompts rather than acting in an agentic fashion. Sora also does not explicitly model physics; all behavior is emergent, which limits control. Finally, Sora is compute-heavy, requiring large-scale compute and data. Sora is the culmination of years of progress in video generation and generative models. It combines advancements in diffusion models, transformer-based reasoning, latent representation learning, and multi-modal conditioning. Sora redefines video models as tools for simulating reality, extending its functionality to define niche areas, such as physics. Ultimately, Sora represents a proxy for world understanding through video [12].

Researchers at Microsoft and Lehigh University provided some additional details about Sora through their testing and evaluation methodologies. From this research, we can determine that the architecture described in this paper details Sora's usage of space-time latent diffusion transformers, which operate over rectangular 3D patches in space and time. This design allows it to model long sequences of video efficiently, producing high-resolution videos spanning up to sixty seconds. Sora leverages a cascade of latent diffusion models: an initial base model generates a low frame-rate video in latent space, and subsequent models refine both spatial resolution and temporal dynamics. The model is trained end-to-end using a combination of video-text pairs, unlabeled videos, and image-caption datasets, helping it ground both appearance and motion in natural language.

A distinguishing feature of Sora is its emergent understanding of physical dynamics and object interactions, even in fictional environments. Sora can simulate body motion, deformation, fluid interactions, and different camera angles without any explicit 3D supervision or integrated physics engine. This behavior suggests that large-scale generative models can internalize aspects of the physical world through predictive training alone. However, Sora's reasoning is purely implicit and statistical, which can sometimes lead to inconsistencies and visual artifacts, especially in edge cases. Despite these limitations, Sora sets a new benchmark in video generation and marks a significant step towards a world model that fuses perception, language, and simulation [13].

B. Open-Sora Plan (2024)

While Sora provided many innovations to the fields of video generation and computer vision, its closed-source nature has made it difficult for the research community to build upon these discoveries. Singapore-based startup HPC-AI Tech, the creators of Colossal-AI, worked to create an open-source model that encapsulates some of the innovations

created by Sora. Open-Sora Plan is an open-source project that aims to contribute a large generation model for generating high-quality videos with long durations. Open-Sora attempts to replicate the video generation functionality and quality offered by OpenAI's SORA model while keeping the model open-sourced. This effort promotes democratized access to advanced video generation techniques and serves as a research platform for further exploration in multi-modal generative modeling.

The model can be broken down into three components: A wavelet-flow variational autoencoder, which obtains multi-scale features in the frequency domain through multi-level wavelet transformation. A joint image-video denoiser, which uses a 3D full attention structure, enhancing the model's ability to understand the world. This makes it capable of creating both high-quality video and images with specific designs. Finally, a condition controller that introduces image conditions into the basic model to support various tasks. Some training strategies are also used to enhance the entire pipeline: Min-max token strategy, which involves min-max tokens for training that aggregate data of different resolutions and durations within the same bucket. Adaptive gradient clipping strategy, which detects outlier data based on the gradient norm at each step and prevents outliers from skewing the model's gradient direction. Finally, a prompt refinement strategy, which enables the model to reasonably expand input prompts while following semantics. This alleviates inconsistencies in prompt length and descriptive granularity, enhancing the stability of video motion and enriching details.

In summary, Open-Sora provides an open-sourced reimplementation of SORA, with all weights and code released to the public. The Wavelet-Flow VAE efficiently compresses spatial-temporal video data, improving the generation quality of the model. Open-Sora also provides a Skiparse Denoising Transformer, a transformer that enables both image and video diffusion, increasing frame consistency. Open-Sora also provides training/sampling strategies and a scalable framework, both of which assist in propelling research in open-source video generation. Open-Sora is computationally intensive, requiring significant resources to function. The model also has a heavy reliance on well-engineered datasets, which are not always optimal to create in an open-source setting. Open-Sora also struggles with consistency in longer videos, and prompts tend to be highly generalized by the system, which may create difficulties when adapting this technology for other domains [14].

C. *Open-Sora 1.2 STDiT (2024)*

Open-Sora Plan was the initial release to replicate some functionality of Sora in an open-source environment. Open-Sora Plan focused on architectural completeness, conditioning, and multi-stage architecture to emphasize a plugin-style design. Open-Sora 1.2 (aka. Open-Sora STDiT), however, focused on performance, modularity, scalability, and flexible video generation across various durations, resolutions, and formats. Open-Sora 1.2's focus is to further democratize efficient, flexible video production.

Open-Sora provides the Spatial-Temporal Diffusion Transformer (STDiT) framework and a 3D compressive autoencoder. The STDiT framework separates the spatial and temporal modeling, improving efficiency and control. The 3D compressive autoencoder speeds up training by compressing latent representations. Limitations include constraints on resolution, which maxes out at 720p across a 15-second duration. Another limitation is the decoupled attention, which could compromise fine-grained temporal coherence in situations with complex motions. The focus of this version is also centered around capability and flexibility rather than saving on compute. Open-Sora 1.2 builds between the two other models mentioned. Compared to Open-Sora Plan, 1.2 improves on the original architectural design by incorporating STDiT and expanding functionality to longer videos with different resolutions. Compared to Open-Sora 2.0, 1.2 focuses on architecture and ecosystem, while 2.0 focuses on consumption and budget [15].

D. *Open-Sora 2.0 (2025)*

While the quality of video generation has improved, it has come at the cost of larger model size, increased data quantity, and greater demand for training compute. Existing models like MovieGen and Step-Video-T2V are expensive to train, requiring millions of dollars in compute. Open-Sora 2.0 attempts to show that top-tier video generation can be done at lower price points, making it more widely available to researchers and developers.

Open-Sora 2.0 provides three major contributions: a cost-efficient training pipeline, a novel video DC-AE autoencoder, and strong performance at low cost. The cost-efficient training pipeline splits training into three stages: text-to-video model on low-resolution video data, image-to-video model on low-resolution video data, and fine-tuning the image-to-video model on high-resolution video. Training cost totaled $199.6k, significantly less than state-of-the-art models. The novel video DC-AE autoencoder reduces token counts by ~5x compared to alternatives. Strong performance at low costs comes as a byproduct of model design, where human evaluations and VBench scores showed that Open-Sora 2.0 can outperform existing models, such as HunyuanVideo and Runway Gen-3, at a reduced budget.

While costs were reduced, the training pipeline still required hundreds of GPUs to process the data. Open-Sora 2.0's usage of latent compression could potentially complicate high-res adaptation. Consistency could also become a limitation as diffusion models suffer from occasional distortion. High-resolution videos are also still a challenge. Open-Sora 2.0 is currently limited to 768x768 resolution across 5-10 seconds [16].

Overall, Sora and Open-Sora created many innovations in video generation and generative AI that are used in most models. Their diffusion-based engines combined with efficient, scalable pipelines for training have created models that provide high-quality, long-duration videos.

IV. Applications of Video Generation Models

There are a variety of applications for video generation currently being researched. These challenges are vast and require domain-specific engineering practices and knowledge to research and innovate in these fields. In this paper, we will discuss applications in optometry, interior design, and mobile computing.

A. *Optometry Simulations*

Fundus fluorescein angiography (FFA) is a technique in optometry that enables the dynamic visualization of retinal blood flow and lesional changes through an injection of

fluorescein sodium. This technique is crucial for assessing certain issues with our eyes, such as blood-retina barrier function, and diagnosing various retinal diseases. Performing FFAs is costly, time-consuming, and requires real patience, which makes it challenging for students and researchers to access consistent, high-quality training material. While diffusion models are already being used in the medical field to assist with a variety of tasks, generating videos with textual descriptions has not been thoroughly explored. To solve this problem, an FFA Sora framework, a custom dataset, domain-conditioned prompts, and an evaluation protocol are provided. The FFA Sora framework is a novel video generation pipeline that uses text prompts, such as descriptions of eye conditions, to synthesize angiography-like videos, inspired by Sora's diffusion architecture. Introduced a small, curated dataset of FFA sequences that are annotated and used to guide the simulation using spatiotemporal patterns. The domain-conditioned prompts combine disease keywords, eye anatomy terms, and phase annotations to guide the generation process. The evaluation protocol uses quantitative metrics and visual assessment to measure realism and fidelity of the generated angiography videos.

Some limitations of this approach include a limited dataset size, which could limit generalizability across patient types. Next, this research does not claim diagnostic utility, which means that while the results are promising, the research is not quite ready to be introduced into real-world clinical situations. Prompts are also templated, so while they provide solid structure, they may not fully leverage clinical notes or other medical documentation. This research shows a promising application of Sora-like video generation tools in the medical domain. It represents the importance of domain conditioning, potential for uses in education and training, and provides a structured model for evaluating quality with a combination of quantitative metrics and human evaluation [17].

B. Interior Design and Architecture

One paper explores the possibility of using a Sora-like model to assist with interior design. While this is conceptual in nature and has not been tested with an actual model, the ideas are promising for future research. The idea is to determine if a Sora-like model could assist interior designers and architects with the early stages of concepting and visualization. Traditional interior design workflows often require time-consuming 3D modeling or physical mockups. Clients may also struggle to visualize spatial ideas described verbally or in 2D sketches. First, this approach presents a conceptual framework for Sora's usage in designs. It's a simple yet practical pipeline for generating videos from design-related text prompts, suggesting how designers might integrate generative models into their workflows. With this framework, a series of prompt engineering techniques are also highlighted, emphasizing descriptors and functionality. Use cases are also demonstrated, such as depicting kitchen layouts, living room designs, and color schemes.

There are some associated limitations with this approach. First, there is no technical implementation or experimentation, which would require more research to understand and benchmark. Use cases are also simplified and do not incorporate the edge cases and ambiguity associated with interior design. Client constraints are also ignored, which is an important component of the interior design process. This research provides conceptual validation of the usage of Sora-like models in a practical creative domain, showing that professionals outside of AI/ML research are considering the usages of generative AI and computer vision and how text-to-video generators can enhance their workflows [18].

C. Mobile Device Video Generation

Another potential application of video generation models is within mobile devices. Research is being done to determine the practical application of large-scale video generation models on mobile and edge devices, where compute and memory constraints make traditional video generation models infeasible. This is especially important for privacy-sensitive, low-connectivity, low-latency applications. This research demonstrates a Sora-style text-to-video generation on an iPhone 15 Pro, using a transformer-based diffusion model truncated and adapted for mobile inference. Token merging and quantization a strategies implemented in this architecture to reduce runtime without sacrificing video quality. This approach allowed inference to fit within 6GB of RAM and generate 5-second videos at 720p quality. Dynamic workload scheduling is another technique used in this approach, which uses a runtime-aware scheduler that balances power consumption and responsiveness by adapting compute intensity to battery level, device temperature, and usage mode. With these components, a benchmarking suite was also released to evaluate video quality, device power consumption, and other metrics.

There are some known limitations with this approach. First, the on-device model is a stripped-down version of Sora, which trades off aspects like visual realism and duration for size and efficiency. Generation is also limited to 5-second videos at 720p resolution. Prompt handling is also narrow and is limited to optimized patterns used during fine-tuning. Finally, using the Apple silicon A17 Pro raises questions about translating this approach to other devices that are potentially less powerful. This research displays an important milestone for translating video generation models to mobile devices. This demonstrates the feasibility of deploying large-scale generative models in real-world applications on mobile devices. It also shows how architecture and system-level optimizations can bridge the gap between modern models and hardware constraints [19].

V. Future Research Directions

While multiple video generation innovations have greatly improved the functionality and efficiency of video generation models, there are still challenges in the field and more opportunities for research. In this paper, we will discuss three potential research paths and their implications: controllable video generation, token compression, and incorporated human feedback.

A. Controllable Video Generation

Current models like Sora and Open-Sora create videos passively from prompts. This can create limitations when users want the ability to have control over dynamic elements. Some research is being explored to extend pretrained video generation models by integrating non-textual conditions, such as camera motion, depth maps, and human pose. A paper by [20] outlines and surveys some developments in this area. The goal is to address the gap between current text-conditioned video generation models and a user's desire to have fine-grained control over their detailed requirements.

This research is to outline some of the popular multi-modal control signals and how they can be best integrated into video generation.

The types of control that are proposed in this approach are structure, ID, image, temporal, audio, and universal. Structure control involves giving users the ability to dictate the spatial layout and configuration of objects. ID control refers to the ability to preserve and manipulate the visual identity of entities within the video. Image control involves user control over video generation from a source image. Temporal control refers to the ability to regulate the evolution of motion and timing across frames. Audio control refers to the ability to synthesize videos from audio signals. Universal control is the ability to flexibly manage and integrate various types of input conditions. There are some potential applications for these suggested control mechanisms. Video completion is one of these applications, where control mechanisms can be used to provide modular, structured guidance to fill in missing areas of a video. Next is video composition, using control mechanisms to insert objects into a video. Next is embodied artificial intelligence, where control mechanisms can be provided to an AI agent, giving the agent more control over the generation parameters. Finally, world modeling in autonomous vehicles, where these control mechanisms can give an autonomous vehicle more control over how it perceives its environment and adapts to its surroundings. These potential applications pose interesting questions about how fine-grained control over the generation process could improve both video generation itself as well as AI and autonomous systems.

While there is more research being done on this topic, there are still a set of challenges that have yet to be solved and potential areas to be researched. Achieving unified control in video generation poses a challenge due to the need to balance multiple user constraints. Another topic is unified video reasoning and generation, where LLMs in video generation systems struggle to handle complex, multi-turn interactions. One area of research to deal with this challenge is Multi-modal LLM agents that can handle inputs across diverse formats and act as a gateway between these formats and the user. Ensuring temporal coherence also remains a challenge as diffusion models tend to struggle with smooth transitions and frame progression, while Autoregressive models tend to struggle with high resource usage for long sequences and high resolutions, limiting scalability. A potential research direction is a hybrid approach that combines a diffusion model and an autoregressive model, where the diffusion model works in latent space to generate representations while the autoregressive model refines them. Overall, further control over video generation could potentially expand the capabilities of video generation models like Sora and Open-Sora, providing avenues for both users and autonomous agents to refine videos, adapt to their environments, and increase the accuracy of the video components [20].

B. Token Compression

As video understanding and analysis tasks are passed to AI agents, performance bottlenecks are beginning to form and deteriorate video understanding, especially videos at or longer than an hour. One factor that appears in these bottlenecks is token compression, specifically the quality of tokens and the compression ratio used. Multi-modal LLM agents have been shown to perform well for video understanding tasks; however, token compression limitations can lead to problems for the AI agents, such as increased memory costs and token limits.

The authors of [21] propose some potential solutions that could lead to further investigation and research. A learnable, self-supervised module called Reconstructive Compression of Tokens (ReCoT) can create compact yet content-rich video tokens. This module uses two major components, a Dynamic Token Synthesizer (DTS) and Semantic-Guided Masking (SGM). DTS incorporates spatiotemporal attention blocks to enhance intra-token relationships, which aids in capturing dynamic movement over time. SGM adaptively masks tokens by considering both intra-frame information density and inter-frame spatiotemporal significance, which facilitates a more efficient self-supervised reconstructive learning process. In conjunction, a dataset pruning strategy is also proposed to filter out redundant or low-informative videos to improve fine-tuning efficiency and better generalization, which mitigates computational and time costs. This approach also proposes Video-XL-Pro, a 3B parameter model specializing in long video understanding. Experiments show that Video-XL-Pro at 3B parameters outperforms most existing 7B parameter models across several benchmarks while also improving model efficiency, processing 8K frames on a single A100 GPU.

Overall, this research poses interesting questions about increasing the effectiveness of multi-modal LLM agents, improving both the quality of their video understanding as well as their efficiency. Further research into token compression strategies and dataset pruning could prove advantageous for future implementations of video generation models [21].

C. Incorporated Human Feedback

While video generation has achieved significant advances through rectified flow techniques, challenges still arise, such as unsmooth motion and misalignment between video and prompts. To align generated videos more closely with human preferences, the authors of [22] propose an approach using human feedback.

A large-scale 182K-sized human-labeled video generation preference dataset, sourced from twelve video generation models, is provided. This dataset is evaluated on three metrics: Visual Quality (VQ), Motion Quality (MQ), and Text Alignment (TA). A validation dataset with 13K annotated triplets is also provided in conjunction with the base dataset. A reward model called VideoReward that learns from human preferences is also provided. This model is grounded in Bradley-Terry pairwise comparison, which focuses on estimating relative preferences between items rather than score regression's approach, which predicts a numerical score for each item based on the comparisons. Finally, three flow-based alignment algorithms are proposed: Flow-DPO, Flow-RWR, and Flow-NRG. Flow-DPO (Direct Preference Optimization) is a training-time RL-based optimization algorithm that incorporates KL regularization. Flow-RWR (Reward Weighted Regression) is a training-time reweighted regression algorithm based on reward signals. Flow-NRG (Noisy Reward Guidance) is an inference-time reward-guided sampling algorithm, allowing users to customize trade-offs between VQ, MQ, and TA.

While the reward dataset is based on manual human annotations, using an automated reward model can eliminate

some costly human effort while also providing videos tailored to user preferences. This research suggests the potential to improve existing methods with human preferences factored into training and provides a good foundation for future research in this area [22].

VI. Conclusion

In this paper, we have outlined milestones in the recent history of image and video generation. Next, we focused on current video generation models utilizing generative AI, such as Sora and Open-Sora's with their text-to-video capabilities. We also highlighted some of Sora and Open-Sora's contributions to the field and technical innovations, such as Sora's usage of space-time latent diffusion transformers and Open-Sora's open-source environment and impressive benchmarking results. We also outlined some potential applications for these technologies, such as interior design and optometry. Finally, we discussed future research directions to expand the field of research, such as token compression and controllable video generation. While video generation technology has rapidly improved over the last 10 years, more advanced functionality is being researched, providing both excitement about the capabilities humans can harness and questions regarding the ethical and moral implications of such advanced models.

Acknowledgment

I would like to thank Dr. Mohammad Alsmirat for giving guidance on formalized research guidelines and topics in generative AI and computer vision. I also appreciate the support of my peers and faculty of the Department of Computer Science and Information Systems at East Texas A&M University.

References

[1] I. J. Goodfellow et al., "Generative Adversarial Networks," arXiv.org, Jun. 10, 2014. https://arxiv.org/abs/1406.2661

[2] P. Isola, J.-Y. Zhu, T. Zhou, and Efros, Alexei A, "Image-to-Image Translation with Conditional Adversarial Networks," arXiv.org, 2016. https://arxiv.org/abs/1611.07004

[3] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural Discrete Representation Learning," arXiv:1711.00937 [cs], May 2018, Available: https://arxiv.org/abs/1711.00937

[4] A. Razavi, A. van den Oord, and O. Vinyals, "Generating Diverse High-Fidelity Images with VQ-VAE-2," arXiv:1906.00446 [cs, stat], Jun. 2019, Available: https://arxiv.org/abs/1906.00446

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," arXiv:2006.11239 [cs, stat], Dec. 2020, Available: https://arxiv.org/abs/2006.11239

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," arXiv:2112.10752 [cs], Apr. 2022, Available: https://arxiv.org/abs/2112.10752

[7] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video Diffusion Models," arXiv:2204.03458 [cs], Jun. 2022, Available: https://arxiv.org/abs/2204.03458

[8] C. Saharia et al., "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," arXiv:2205.11487 [cs], May 2022, Available: https://arxiv.org/abs/2205.11487

[9] U. Singer et al., "Make-A-Video: Text-to-Video Generation without Text-Video Data," arXiv:2209.14792 [cs], Sep. 2022, Available: https://arxiv.org/abs/2209.14792

[10] H. Chen et al., "VideoCrafter1: Open Diffusion Models for High-Quality Video Generation," arXiv.org, 2023. https://arxiv.org/abs/2310.19512

[11] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and Content-Guided Video Synthesis with Diffusion Models." Available: https://openaccess.thecvf.com/content/ICCV2023/papers/Esser_Structure_and_Content-Guided_Video_Synthesis_with_Diffusion_Models_ICCV_2023_paper.pdf

[12] OpenAI, "Video generation models as world simulators," Openai.com, 2024. https://openai.com/index/video-generation-models-as-world-simulators/

[13] Y. Liu et al., "Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models," arXiv.org, Feb. 28, 2024. https://arxiv.org/abs/2402.17177

[14] B. Lin et al., "Open-Sora Plan: Open-Source Large Video Generation Model," arXiv.org, 2024. https://arxiv.org/abs/2412.00131

[15] Z. Zheng et al., "Open-Sora: Democratizing Efficient Video Production for All," arXiv.org, 2024. https://arxiv.org/abs/2412.20404

[16] X. Peng et al., "Open-Sora 2.0: Training a Commercial-Level Video Generation Model in $200k," arXiv.org, 2025. https://arxiv.org/abs/2503.09642v1 (accessed Jul. 22, 2025).

[17] X. Wu et al., "FFA Sora, video generation as fundus fluorescein angiography simulator," arXiv.org, 2024. https://arxiv.org/abs/2412.17346 (accessed Jul. 25, 2025).

[18] Dr and Virendra Gawande, "Enhancing Interior Design with OpenAI Sora: A Fusion of Creativity and Computational Intelligence," International Journal of Engineering Research and Applications www.ijera.com, vol. 14, pp. 56–61, 2024, doi: https://doi.org/10.9790/9622-14055661.

[19] B. Kim, K. Lee, I. Jeong, J. Cheon, Y. Lee, and S. Lee, "On-device Sora: Enabling Training-Free Diffusion-based Text-to-Video Generation for Mobile Devices," arXiv.org, 2025. https://arxiv.org/abs/2502.04363 (accessed Jul. 25, 2025).

[20] Y. Ma et al., "Controllable Video Generation: A Survey," arXiv.org, 2025. https://www.arxiv.org/abs/2507.16869? (accessed Jul. 27, 2025).

[21] X. Liu, Y. Shu, Z. Liu, A. Li, Y. Tian, and B. Zhao, "Video-XL-Pro: Reconstructive Token Compression for Extremely Long Video Understanding," arXiv.org, 2025. https://arxiv.org/abs/2503.18478 (accessed Jul. 27, 2025).

[22] J. Liu et al., "Improving Video Generation with Human Feedback," arXiv.org, 2025. https://arxiv.org/abs/2501.13918 (accessed Jul. 27, 2025).