Generate Impressive Videos with Text Instructions: A Review of OpenAI Sora, Stable Diffusion, Lumiere and Comparable Models

Enis Karaarslan¹ and Ömer Aydın¹

¹Affiliation not available

March 19, 2024

Abstract

The generation of videos from textual input poses a significant computational challenge within computer science. Nonetheless, recent advancements in text-to-video artificial intelligence (AI) technologies have showcased notable progress within this domain. Foreseen advancements in realistic video generation and data-driven physics simulations are poised to further propel the field forward. The emergence of text-to-video AI holds transformative potential across a plethora of creative domains, including filmmaking, advertising, graphic design, and game development, as well as within sectors such as social media, influencer marketing, and educational technology. This research study seeks to comprehensively review generative AI methodologies in text-to-video synthesis, with an emphasis on large language models and AI architectures. Multiple methods such as literature review, technical evaluation and solution proposal were applied for this purpose. Prominent models such as OpenAI Sora, Stable Diffusion, and Lumiere are evaluated for their efficacy and architectural intricacies. However, the pursuit of Artificial General Intelligence (AGI) is accompanied by a myriad of challenges. These encompass the imperative to safeguard human rights, prevent potential misuse, and protect intellectual property rights. Ensuring the accuracy and integrity of the generated content is paramount. The computationally intensive nature of transformer models results in substantial electricity and water consumption, necessitating the formulation of strategies to mitigate environmental and computational costs to ensure long-term sustainability. This research endeavors to explore potential avenues for addressing these challenges and proposes solutions to advance environmental and computational efficiency within the context of text-to-video AI.
Generate Impressive Videos with Text Instructions: A Review of OpenAI Sora, Stable Diffusion, Lumiere and Comparable Models

Enis Karaarslan*, Ömer Aydın
*enis.karaarslan@mu.edu.tr

Abstract
The generation of videos from textual input poses a significant computational challenge within computer science. Nonetheless, recent advancements in text-to-video artificial intelligence (AI) technologies have showcased notable progress within this domain. Foreseen advancements in realistic video generation and data-driven physics simulations are poised to further propel the field forward. The emergence of text-to-video AI holds transformative potential across a plethora of creative domains, including filmmaking, advertising, graphic design, and game development, as well as within sectors such as social media, influencer marketing, and educational technology. This research study seeks to comprehensively review generative AI methodologies in text-to-video synthesis, with an emphasis on large language models and AI architectures. Multiple methods such as literature review, technical evaluation and solution proposal were applied for this purpose. Prominent models such as OpenAI Sora, Stable Diffusion, and Lumiere are evaluated for their efficacy and architectural intricacies. However, the pursuit of Artificial General Intelligence (AGI) is accompanied by a myriad of challenges. These encompass the imperative to safeguard human rights, prevent potential misuse, and protect intellectual property rights. Ensuring the accuracy and integrity of the generated content is paramount. The computationally intensive nature of transformer models results in substantial electricity and water consumption, necessitating the formulation of strategies to mitigate environmental and computational costs to ensure long-term sustainability. This research endeavors to explore potential avenues for addressing these challenges and proposes solutions to advance environmental and computational efficiency within the context of text-to-video AI.

Keywords: Generative Artificial Intelligence, Large Language Model, text-to-video AI, transformer, diffusion model, data-driven physics engine

1. INTRODUCTION
The advancement and evolution of technical knowledge are propelled by the digitalization of various processes, with particular emphasis on the remarkable progress observed in artificial intelligence (AI) in recent years. This technological evolution is being further
propelled by advancements in hardware capacity, processing capabilities of computer systems, and the proliferation of cloud computing and edge computing paradigms. Within the realm of AI, the efficacy and applicability of AI models have significantly improved owing to enhanced computational capabilities and the refinement of AI algorithms. Consequently, AI finds widespread application across diverse domains.

In this context, both OpenAI and other competitive entities have introduced their own AI models and tools, with OpenAI gaining global recognition following the release of its ChatGPT chatbot in November 2022. Leveraging this momentum, OpenAI subsequently augmented its AI models with image processing capabilities through the introduction of version GPT-4. OpenAI’s future endeavors in this domain remain high which continues to the creation of videos from the text. While such advancements necessitate substantial computational resources, the prevailing technological landscape suggests that achieving the requisite processing capacities should be within reach.

The main objective of this study is to explore the potential applications enabled by the text-to-video capabilities of Generative AI tools. In this regard, we also seek answers to the following research questions:

- What are the technical evaluations of prominent text-to-video synthesis models, and how do they address the challenges associated with realistic video generation?
- What strategies can be implemented to ensure the trustworthiness and traceability of generated content in text-to-video synthesis?
- What environmental and computational challenges arise from the significant electricity and water consumption associated with transformer models, and how can strategies be formulated to mitigate these costs for long-term sustainability?

2. FUNDAMENTALS

In this section, we will initiate the discussion by elucidating essential concepts pertaining to large language models (LLMs), which serve as the principal catalysts driving modern chatbot technology, progressively integrating within the realm of artificial intelligence. Additionally, we will explore the burgeoning capabilities of video generation within this context.
2.1. Large Language Models

Large Language Models (LLMs) represent a category of artificial intelligence systems adept at processing and generating natural language texts. Leveraging deep neural networks, these systems undergo extensive training on vast repositories of textual data encompassing books, articles, web content, and social media discourse, among other sources. Through this process, LLMs acquire an understanding of the intricate patterns and structures inherent in natural language, enabling them to execute diverse tasks including question answering, text summarization, language translation, and essay composition.

However, despite their formidable capabilities, LLMs exhibit certain limitations that warrant attention. Concerns are raised regarding potential biases ingrained within the training data, the presence of factual inaccuracies, and the absence of ethical considerations in their design and operation. Furthermore, the opaque nature of LLM functioning impedes their explainability and transparency, rendering it challenging to discern the rationale behind their generated responses and operational methodologies.

2.2. Text-to-Image AI

Images are elements that can often be explained with text but require special skill and effort to create. Illustrations, photographs and paintings are examples of this. Therefore, a tool that can produce realistic images using natural language can provide people with unprecedented ease in creating unique visual content. The ability to edit images with natural language offers a significant advantage for real-world applications that are both critical for iterative improvements and support detailed control.

The process of text-to-image (T2I) has been extensively explored in the literature, as evident in several studies (Ramesh et al., 2021, Yu et al.;2021, Yu et al., 2022; Ding et al.,2022; Gafni et al.,2022). Different models for text-to-image generation, such as GLIDE (Nichol et al., 2021), DALLE-2 (Ramesh et al., 2022), and Imagen (Saharia et al., 2022), have been developed based on diffusion models (Ho et al., 2020). GLIDE utilizes CLIP guidance at a scale of 2.0 and classifier-free guidance(Ho and Salimans, 2022) at a scale of 3.0, without performing CLIP reranking or cherry-picking. DALLE-2 enhances text-image alignments by leveraging the CLIP (Radford et al., 2021) feature space. Imagen employs cascading diffusion models for generating high-resolution videos, and subsequent models
like VQ-diffusion (Gu et al., 2022) and Latent Diffusion Models (LDMs) (Rombach et al. 2022) operate in the latent space of an autoencoder to enhance training efficiency. Our approach extends the LDMs by expanding the 2D model into the spatiotemporal domain within the latent space.

2.3. Evolution of Text-To-Video AI

Creating video from the text was a difficult process in the past with computational resource constraints. This often resulted in restricting resolution and complexity. However, this tends to decrease with the development of hardware and software solutions. We can summarize the models and their problems in Table 1 (Martin, 2024).

<table>
<thead>
<tr>
<th>Models</th>
<th>Period</th>
<th>Examples</th>
<th>Working Principle</th>
<th>Disadvantages</th>
</tr>
</thead>
</table>
| Generative Adversarial Networks (GAN) and Variational Autoencoder (VAE) | Mid 2010    | Experimental              | - An iterative process of enhancing the output of a generator through iterative refinement while a discriminator learns to discern between authentic and synthesized videos.  
                             |              |                           | - VAEs encode videos into a latent space and subsequently decoding them back into the video domain, using conditioning based on textual input. | - Challenges in preserving temporal uniformity and coherence across extended video sequences |
| Transformer based models               | Late 2010s  | Phenaki Make-A-Vide o     | Develop sophisticated methods to capture intricate relationships within video frames, enabling the creation of cohesive sequences based on textual descriptions. | - Challenges in producing high-quality content  
                             |              |                           |                                                                                  | - Constraints in achieving diversity and realism in the generated videos. |
| Diffusion models                       | Early 2020s | Imagen Video Make-A-Vide o Stable Diffusion | An iterative procedure of multiple stages refining the initial random pattern, gradually reducing noise, and structuring it into a cohesive video sequence. | - Challenges on producing extended video sequences while ensuring the preservation of narrative coherence throughout their entirety. |
| Spacetime Transformer                 | 2023-2024   | Lumiere Sora              | - Segment videos into smaller units or "patches", Similar to the process of tokenizing text in language models. | - Difficulties arise in effectively modeling intricate interactions and preserving coherence across extended timeframes.  
                             |              |                           |                                                                                  | - Unexpected occurrences of object appearances |

We can summarize the studies as follows. In their study, Lee et al. (2018) trained a conditional generative model to extract both static and dynamic information from the text. They presented a hybrid framework that combines a Variational Autoencoder (VAE) and a Generative Adversarial Network (GAN). The framework utilizes static features ("gist") to
represent text-conditioned background color and object layout structure. In contrast, dynamic features are incorporated by transforming input text into an image filter. To train the deep-learning model, a method is developed to automatically create a matched text-video corpus from publicly available online videos. Experimental results demonstrate that the proposed framework generates realistic and diverse short-duration videos that accurately convey input text information. The framework's performance surpasses baseline models that directly adapt text-to-image generation procedures for video production, as assessed visually and through the inception score commonly used to evaluate image generation in GANs (Le et al., 2018).

Balaji et al. (2019) discuss developing conditional generative models for text-to-video synthesis, a challenging yet crucial aspect of machine learning research. They introduce the Text-Filter conditioning Generative Adversarial Network (TFGAN), a conditional GAN model incorporating a novel multi-scale text-conditioning scheme to enhance text-video associations. The authors also create a synthetic dataset of text-conditioned moving shapes for systematic evaluation. Extensive experiments revealed that TFGAN outperforms existing approaches significantly, showcasing its ability to generate videos of novel categories not encountered during training.

Yan et al. (2021) introduce VideoGPT, a straightforward architecture for the generative modeling of natural videos. VideoGPT employs VQ-VAE, utilizing 3D convolutions and axial self-attention to learn downsampled discrete latent representations of raw videos. A GPT-like architecture is then used to model these discrete latents in a spatiotemporal manner. Despite its simplicity, VideoGPT achieves competitive results with state-of-the-art GAN models in video generation on the BAIR Robot dataset. It also demonstrates the capability to generate high-fidelity natural videos from UCF-101 and the Tumbler GIF Dataset (TGIF). The authors provide samples and code for reproducibility, aiming to offer a minimalistic implementation of transformer-based video generation models (Yan et al., 2021). You can use the VideoGPT from the following Google Colab link (https://colab.research.google.com/github/wilson1yan/VideoGPT/blob/master/notebooks/Using_VideoGPT.ipynb).
Liu et al. (2023) introduce "Generative Disco," a generative AI system designed for text-to-video generation to enhance music visualization. Acknowledging the complexity and resource-intensiveness of traditional music visualization creation, the system utilizes large language models to help users visualize music intervals. It achieves this by identifying prompts to describe the images at interval starts and ends, then interpolating between them to synchronize with the music beat. The authors propose design patterns, namely transitions (for shifts in color, time, subject, or style) and hold (to focus on subjects), to enhance the quality of the generated videos. A study with professionals demonstrates that these patterns provide a highly expressive framework, enabling the creation of coherent visual narratives. Their paper concludes by emphasizing the generalizability of these design patterns and the potential of generated videos for creative professionals.

3. METHOD

More than one method was employed in this study. While it primarily serves as a literature review, it additionally conducts a technical evaluation of specific text-to-video generation models, including OpenAI Sora, Stable Diffusion, and Lumiere. This assessment centers on elucidating the strengths and weaknesses inherent in these models. Subsequently, a comparative analysis was conducted, where select artificial intelligence models employed in video production from the text were scrutinized, and an in-depth analysis of their architecture, performance, and technical intricacies was undertaken to facilitate a comprehensive comparison.

In conclusion, as a solution proposal study, the research puts forth solutions addressing ethical and environmental challenges that arise in the pursuit of artificial general intelligence (AGI). This signifies an approach that is geared towards ensuring the future sustainability of text-to-video production technology from a broader perspective.

4. EVALUATION OF MAIN ACTORS

4.1 Lumiere

Lumiere is a text-to-video diffusion model aimed at creating realistic and coherent motion in synthesized videos, addressing a key challenge in video synthesis. The model employs a
Space-Time U-Net architecture, enabling the generation of the entire temporal duration of the video in a single pass. This is a departure from existing models that use distant keyframes and temporal super-resolution, making global temporal consistency challenging. Lumiere utilizes spatial and temporal down- and up-sampling, along with a pre-trained text-to-image diffusion model, to generate a full-frame-rate, low-resolution video across multiple space-time scales. The model achieves state-of-the-art text-to-video generation results, demonstrating its versatility for various content creation tasks and video editing applications, such as image-to-video, video inpainting, and stylized generation (Lumiere, 2024).

Figure 1. Space-Time UNet (STUNet) architecture (Bar-Tal et al., 2024; Lumiere, 2024)

Bar-Tal et al. (2024) expanded a pre-trained Text-to-Image (T2I) U-Net architecture (Ho et al., 2022) into a Space-Time UNet (STUNet). This STUNet performs both spatial and temporal down-sample and up-sample of videos. Figure 1.a illustrates the activation maps of STUNet, with color indicating features from different temporal modules. In Figure 1.b, convolution-based blocks are shown, consisting of pre-trained T2I layers followed by factorized space-time convolution. Figure 1.c displays attention-based blocks at the coarsest U-Net level, where pre-trained T2I layers are followed by temporal attention (Bar-Tal et al., 2024;).

4.2 Stable Diffusion

Stable Diffusion (Rombach et al., 2022) is an influential class of generative models in the field of image generation from text. The outstanding advantage of Stable Diffusion is its
ability to produce high-quality images. This capability is based on improving throughput by using multiple diffusion steps, increasing control by gradually adding noise and facilitating image synthesis. Continuous improvements have been made in Stable Diffusion version 1.1 since 2022. Using this version, one can better evaluate its ability to create realistic and subject-specific images by analyzing the model's subject-specific training results on custom datasets. Moreover, LMU Munich's Computer Vision and Learning department provides easily accessible Hugging Face compatible formats for Stable Diffusion version 1.1.

Hilado (2023) achieved promising results in his study by training Stable Diffusion 1.1 specifically with the Pokémon dataset. Common issues in pre-trained images, such as redundant text and inconsistencies like unrecognizable characters and changing styles, are evident. This situation is illustrated in Figure 2. Many images generated using Hilado's fine-tuned model more closely align with the features present in subject-specific datasets. Examples of this alignment include white backgrounds, consistent character styles, and color usage. In the first prompt, the fine-tuned model exclusively produced red and white images associated with the characters, whereas the pre-trained model included a character without red. The second prompt provided additional instances of white backgrounds and consistent coloring in the images of the fine-tuned model. In the final prompt, the pre-trained model generated images that diverged from the style of Pokémon characters compared to the fine-tuned model (Hilado, 2023).
OpenAI Sora is an artificial intelligence model that helps people solve real-world problems. This model was developed to teach artificial intelligence to understand and simulate the physical world in motion. Sora can create videos up to one minute long while maintaining visual quality and adherence to the user’s wishes. Sora, which is presented to red team members to evaluate especially critical areas, improves its model by receiving feedback from creative professionals. Every stage of research progress is shared. In this way, it informs people about artificial intelligence capabilities. Sora is a model who can accurately render complex scenes and has a deep understanding of languages to understand prompts and create emotional characters. However, the model also has some weaknesses; may have difficulty physically simulating complex scenes and may have difficulty understanding

<table>
<thead>
<tr>
<th>Prompt</th>
<th>Pre-trained</th>
<th>Fine-tuned</th>
</tr>
</thead>
<tbody>
<tr>
<td>A pokemon red and white cartoon ball with an angry look on its face</td>
<td><img src="sora.png" alt="Image" /></td>
<td><img src="sora.png" alt="Image" /></td>
</tr>
<tr>
<td>A blue dragon pokemon flying through the air</td>
<td><img src="sora.png" alt="Image" /></td>
<td><img src="sora.png" alt="Image" /></td>
</tr>
<tr>
<td>A green pokemon fish with big eyes</td>
<td><img src="sora.png" alt="Image" /></td>
<td><img src="sora.png" alt="Image" /></td>
</tr>
</tbody>
</table>

**Figure 2.** Stable Diffusion 1.1 trained on Pokémon (Hilado, 2023)

**4.3 OpenAI Sora**

OpenAI Sora is an artificial intelligence model that helps people solve real-world problems. This model was developed to teach artificial intelligence to understand and simulate the physical world in motion. Sora can create videos up to one minute long while maintaining visual quality and adherence to the user’s wishes. Sora, which is presented to red team members to evaluate especially critical areas, improves its model by receiving feedback from creative professionals. Every stage of research progress is shared. In this way, it informs people about artificial intelligence capabilities. Sora is a model who can accurately render complex scenes and has a deep understanding of languages to understand prompts and create emotional characters. However, the model also has some weaknesses; may have difficulty physically simulating complex scenes and may have difficulty understanding...
certain cause-effect relationships (Sora, 2024). Sora is a diffusion model that starts with static noise and it creates a video by gradually eliminating that noise. It can create all videos at once or expand existing videos. By using the prediction of many frames at once, the model solves the challenging problem of ensuring that an object remains the same even though it temporarily moves out of view. Sora, which can achieve superior scaling performance thanks to its transformer architecture, uses a structure similar to GPT models. Videos and images are represented as collections of small units of data called patches, making it possible to train on a wider range of visual data. Sora is based on the DALL-E and GPT models and is capable of following text instructions more faithfully using DALL-E 3’s titling technique. In addition to creating a video from text instructions, the model can create a video from an existing image, animate the image, and expand an existing video or fill in missing frames. Sora performs a fundamental task of understanding and simulating the real world.

Sora will undergo various security measures before being used for OpenAI products. While the model will be tested with tools that can detect misleading content, it will be tested against security challenges by working with red team members. The inclusion of C2PA metadata (https://c2pa.org/) is also planned for a future OpenAI product. Security methods used in DALL-E 3 will be leveraged, text classifiers will be used to check for text entries that violate usage policies, and powerful image classifiers will review videos to check whether they comply with usage policies. By communicating with policymakers, educators, and artists, concerns will be understood and positive examples of the use of the technology will be identified. However, it may not be possible to predict all positive and negative usage scenarios, so the importance of the process of learning from real-world usage is emphasized (Sora, 2024).

The article titled “Video generation models as world simulators” (OpenAI, 2024) co-authored by OpenAI researchers, provides significant insights into Sora. Within this article, it is highlighted that Sora possesses the capability to produce videos with diverse aspect ratios and resolutions. The article outlines Sora's versatility in executing various image and video editing tasks, ranging from crafting looping videos to adjusting temporal aspects like forward or backward stretching. Additionally, Sora can alter the background of
an existing video. A particularly intriguing revelation in the article pertains to Sora's capacity to "simulate digital worlds." Notably, an experiment detailed in the article involved prompting Sora with the term "Minecraft," resulting in Sora adeptly generating a Minecraft-like Heads-Up Display (HUD), gameplay, and even replicating the game's dynamics, including physics. This experiment showcased Sora's ability to simultaneously create and control the player character within the simulated digital environment (Wiggers, 2024).

Achieving consistency and repeatability across multiple scenes in AI-generated videos was previously a significant challenge. This difficulty arose from the fact that, when creating each scene or frame individually, it was challenging to fully understand previous contexts and details and to consistently incorporate them into the subsequent scene. However, by integrating a comprehensive understanding of language with the visual context and accurately interpreting instructions, this model is capable of maintaining story coherence. Additionally, it can effectively portray the characters' personalities and emotions in line with the instructions, making them stand out as remarkable characters in the film. Sora expands upon earlier research on models for generating image data. Earlier studies have employed a variety of techniques, including diffusion models, autoregressive transformers, recurrent networks, and Generative Adversarial Networks (GANs). However, these techniques are often limited to a certain type of visual data, such as fixed-size or brief movies. To get beyond these restrictions and produce films with varying lengths, aspect ratios, and resolutions, Sora has undergone substantial development (Kiuchi, 2024). We will share the main technologies that underpin these advancements in the following subsections.

4.3.1 Data driven physics engine

The application of physics to games has evolved significantly, departing from the original practice of programming special effects into individual games. Originally, each game would have its physics programmed to meet its unique requirements. For example, if a game needs arrows to follow trajectories, the trajectory equation must be programmed specifically for that game. While suitable for simple simulations, this approach became less effective as the game's complexity increased. As games required more realistic physics and similar effects were needed in different games, developers looked for generic solutions that could be reused. This led to the development of "physics engines", common code that can
simulate general physics without depending on the specifics of each game's scenario. However, a difficulty arose when dealing with certain objects such as arrows or bullets. The physics engine functions as a mathematical calculator to simulate physics but lacks information about what should be simulated. Besides the engine, game-specific data representing the game level is also important (Millington, 2007).

Data are the raw material of applied statistics - providing a focal point or pivot in statistical methodology. Data-driven methods refer to approaches in which the parameters or behavior of the model are determined by existing data rather than being determined by explicitly programmed rules. These methods leverage large datasets to identify patterns, correlations, and relationships within the data. This allows the model to make predictions, classifications or decisions based on the learned information (Sprent, 2019).

When comparing physics-based solutions and data-driven approaches, we can conclude that neither is inherently superior to the other. The complexity of the systems we attempt to model today has become so uncertain that a pure physics-based approach often reaches a dead end. Even if it doesn't lead to an impasse, the intricate nature of the relevant calculations remains high, rendering the time required for computation less effective and potentially impacting the environment. Self-learning models, based on real-world test data, have demonstrated an ability to surpass what can be achieved with physics-based models alone (Murad, 2023).

4.3.2 Transformers

This model can be described as a revolutionary neural network architecture in the field of natural language processing (NLP), as proposed by Vaswani et al. (Vaswani et al., 2017). The model overview can be seen in Figure 3. It significantly overcomes the challenges encountered in traditional recurrent neural networks (RNNs) and convolutional neural networks (CNNs).
Issues with RNNs include the challenge of capturing long-term dependencies and limitations on parallelization, making training on large datasets inefficient. On the other hand, CNNs face problems with fixed receptive field size and struggle to model the hierarchical structure of natural language. The transformer introduces novel features such as the attention mechanism, enabling direct modeling of dependencies across any positions in the sequence, and addressing the issue of capturing long dependencies. Additionally, the realization of parallelization in computation significantly accelerates training on large datasets. The variable receptive field, facilitated by the attention mechanism, allows the model to dynamically adjust its focus on local information for specific tasks while considering broader context in other cases. These features collectively make the Transformer a groundbreaking architecture for natural language processing (Kiuchi, 2024).

4.3.3 Vision Transformer (ViT)
By examining token relations, the original transformer paper proposed the idea of using tokens to represent individual words or phrases, allowing for a deeper comprehension of sentence meaning. This study expands this concept to visual data by splitting images into 16x16 patches and considering each patch as a "token" within the transformer. The fixed receptive field size constraints of conventional CNN models in image recognition are overcome and the model can learn associations between patches by using this technique. This allows for the variable capturing of positional relationships inside a picture, which improves the model's overall visual context recognition and comprehension (Kiuchi, 2024). The model proposed by Dosovitskiy, et al. (2020) is given in Figure 4.

**Figure 4:** Vision Transformer model overview (Dosovitskiy, et al., 2020).

### 4.3.4 Video Vision Transformer (ViViT)

Expanding upon the vision transformer concept, ViViT takes it a step further by applying it to the intricate realm of videos, which involves both static spatial elements and dynamic temporal elements. To address the complexity of video data, ViViT breaks down videos into spatiotemporal patches, treating them as tokens within the Transformer model. This
innovative approach allows ViViT to concurrently grasp static and dynamic elements within videos, effectively modeling the intricate relationships between them (Kiuchi, 2024). An alternative approach, illustrated in Figure 5, involves extracting non-overlapping spatio-temporal "tubes" from the input volume and linearly projecting them into $\mathbb{R}^d$. This technique expands ViT's embedding to three dimensions, essentially representing a 3D convolution. In essence, this method integrates spatio-temporal information during tokenization, differing from the "Uniform frame sampling" method where the transformer fuses temporal information from various frames. Non-overlapping tubelets spanning the spatio-temporal input volume are extracted and linearly embedded (Arnab et al., 2021).

![Figure 5: Tubelet embedding (Arnab et al., 2021).](image)

### 4.3.5 Masked Autoencoders (MAE)

He et al. (2022) significantly addressed the historically high computational costs and inefficiencies in training on extensive datasets linked with high dimensionality and vast information. Employing the self-supervised pre-training method, masked autoencoder, the study improved efficiency by training the network to predict hidden information in the input
image through masking. This led to a more effective learning process for crucial features within the image, resulting in rich visual data representations. Consequently, this approach enhanced data compression, reduced computational costs, and increased adaptability across various types of visual data and tasks. The study's methodology is closely connected to advancements in language models, like BERT, which achieved deep contextual understanding in text through Masked Language Modeling (MLM). Similarly, the study applied a comparable masking technique to visual data, contributing to a more profound understanding and representation of images. Masked autoencoders architecture addressed by He et al. (2022), can be seen in Figure 6.

![Masked Autoencoders Architecture](imageurl)

**Figure 6:** Masked Autoencoders Architecture (He et al., 2022).

### 4.3.6 Native Resolution Vision Transformer (NaViT)

Dehghani et al. (2024) proposed the Native Resolution ViTransformer (NaViT), a model designed to further expand the applicability of the vision transformer (ViT) to images of any aspect ratio or resolution. This study challenges the common practice of resizing images to a fixed resolution before processing with computer vision models. Introducing NaViT, the
approach utilizes the flexible sequence-based modeling of ViT, allowing for variable input sequence lengths. NaViT employs sequence packing during training to handle inputs with arbitrary resolutions and aspect ratios, demonstrating improved training efficiency for large-scale supervised and contrastive image-text pretraining.

**Figure 7.** Native Resolution Vision Transformer (NaViT) (Dehghani et al., 2024)

The model can be efficiently applied to standard tasks like image and video classification, object detection, and semantic segmentation, resulting in enhanced performance on robustness and fairness benchmarks. The input resolution flexibility of NaViT enables a smooth navigation of the test-time cost-performance trade-off during inference. This departure from the conventional CNN-designed input and modeling pipeline suggests a promising direction for ViTs (Dehghani et al., 2024). Dehghani et al. (2024) present an alternative, NaViT shown in Figure 7. Multiple patches from different images are packed in
a single sequence—termed Patch n’ Pack—which enables variable resolution while preserving the aspect ratio.

4.3.7 Diffusion Models

Sora relies on Diffusion Models in conjunction with the Transformer as its fundamental technology. The theoretical groundwork for diffusion models, based on non-equilibrium thermodynamics, was laid by Sohl-Dickstein et al. in 2015. These models introduce the concept of a diffusion process, starting with random noise and systematically removing it to generate data resembling real images or videos. This methodology, originally envisioned for creating complex data from random dots, has been successfully employed in generating high-quality images and sounds, contributing to the advancement of sophisticated generative models (Kiuchi, 2024).

4.3.8. Latent Diffusion Models

Rombach, et al. (2022) significantly advances high-resolution image synthesis using diffusion models. It introduces a method that reduces computational costs by employing diffusion models in the latent space, achieving comparable results with fewer resources. For Sora, this approach extends to video data, efficiently processing information in a lower-dimensional latent space and decomposing it into spatiotemporal patches. This enables Sora to rapidly generate higher-quality visual content.

4.3.9 Diffusion Transformer (DiT)

Peebles and Xie's (2023) research might be the most crucial in realizing Sora. As mentioned in the technical report published by OpenAI (OpenAI, 2024), Sora employs not a vanilla (normal) transformer but a diffusion transformer (DiT). The study presented a novel model that substitutes the U-net component in diffusion models with a transformer structure, forming the Latent Diffusion Model. This structure operates on latent patches, enhancing the efficiency of image patch handling and enabling the generation of high-quality images with optimized computational resources. Different from Stability AI's Stable Diffusion, the inclusion of this Transformer is believed to enhance the generation of more natural videos (Kiuchi, 2024).
5. RESULTS

The video titled "Pepperoni Hug Spot - AI Made TV Commercial" (https://www.youtube.com/watch?v=qSewd6laj6I) is an experimental fiction on shooting commercial videos with artificial intelligence and represents one of the initial attempts at advancements in this field. The screenshot of the video, generated by the Runway AI platform in April 2023, is presented in Figure 8. Upon a cursory examination, the video exhibits semantic integrity, but upon closer inspection, significant distortions and discrepancies with reality become apparent to the viewer. The video portrays a family comprising a mother, a father, a girl, and a boy enjoying pizza. While this scene may seem ordinary at first glance, a detailed analysis reveals distortions and deformities in the facial features and limbs of the individuals, clearly indicating that the video was not captured in the real world but rather created by a computer.

![Figure 8. Screenshot from the experimental commercial video “Pepperoni Hug Spot”](image)

Furthermore, the video shared at https://openai.com/sora, depicted in Figure 9, is generated by the OpenAI Sora using the text from Prompt 1. This video was released nine months after the Runway AI platform's video in April 2023. Distinguishing a freshly generated AI video from one shot in the real world proves challenging, especially for non-experts. Prompt 1’s text comprises six sentences that vividly describe the environment, accessories, clothing, and visual characteristics of the woman in the video, as well as the
street and other people around her. OpenAI Sora successfully generates a 59-second video with high fidelity, making it difficult for non-experts to discern its artificial origin.

However, individuals aware of the AI-generated nature may identify minor logic and reality issues, particularly in the limbs of the woman and background people, albeit subtle. Experts in the field can easily detect these anomalies. Despite a mere nine-month gap between the two examples, a noticeable and satisfactory acceleration in development is evident, manifesting as improved quality and realism in the videos. As development continues, it is conceivable that even more realistic and extended AI-generated videos will soon become commonplace.

**Prompt 1:** “A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.”

![Figure 9. 59-second video generated by OpenAI Sora using Prompt 1 (Sora, 2024)](image)

6. **RISK AND CONSTRAINTS**

The primary concerns to the deployment of transformers are computational and environmental in nature. It is widely acknowledged that the utilization of transformers entails
considerable computational overhead and places substantial demands on GPU (Graphics Processing Unit) capabilities. Elon Musk, a prominent figure in the technological sphere, has underscored the potential ramifications of such demands on the global electricity infrastructure. In a keynote address delivered during the closing question and answer session at the Bosch Connected World conference in 2024, Musk articulated concerns regarding the availability of requisite electrical infrastructure, particularly transformers, for the sustained operation of transformative technologies. Specifically, he cautioned against impending shortages in transformer availability, attributing this to the escalating deployment of Artificial Intelligence (AI) systems and the burgeoning adoption of electric vehicles (EVs), both of which exert significant strain on global energy resources. Musk's remarks signal the imperative for proactive measures to address the anticipated challenges in electrical infrastructure provisioning to support the accelerating pace of technological innovation.

The water consumption of AI systems is mostly overlooked. Water consumption in operational processes encompasses onsite and offsite dimensions. Onsite usage relates to water utilized within the operational confines of a facility, such as data centers, where significant volumes are employed for cooling purposes through mechanisms like cooling towers or external air intake. Conversely, offsite consumption encompasses water utilization in electricity generation that includes substantial usage in the cooling systems of power plants, as well as water evaporation in hydropower operations. Moreover, the supply chains of AI technologies entail embodied water consumption, evident in microchip production requiring considerable amounts of Ultra-Pure Water (UPW). Additionally, large language models contribute to freshwater consumption during training phases and query executions. Transformer operations will also require more water as well. The water footprint can be calculated with Formula 1. Here, the onsite Water Usage Effectiveness (WUE) quantifies the efficiency of water usage in cooling systems. The Power Usage Effectiveness (PUE) assesses the energy overheads apart from IT operations, encompassing cooling energy and power distribution losses. Additionally, the offsite WUE, also known as the electricity water intensity factor, evaluates the efficiency of water usage in electricity generation processes (Ren, 2024).

WaterFootprint = ServerEnergy x WUE(Onsite) + ServerEnergy x PUE x WUE(Offsite) (1)
Watermarking like C2PA has also constraints; this metadata can easily be removed from the images especially when transferred to another system like social media (David, 2024).

7. DISCUSSION

Most of the products are limited to generating short (around 4-second) clips, but Sora's capacity to produce 60-second videos establishes a novel benchmark within the industry. People tend to have a fear of too realistic human-like robots (Mori, MacDorman & Kageki, 2012) and avatars (Altundas & Karaarslan, 2023) as well, and this is called the uncanny valley. Most of the videos do not contain realistic human content and some professionals evaluated these as the uncanny valley. Sora demonstrates exceptional clarity, fluidity of motion, and remarkable fidelity to human anatomy and real-world physical attributes. However, Sora will also face serious competition from various products, including startups like Runway Gen-2, Pika Labs, Stability AI and Lumiere from Google.

These improvements are a threat to stock video agencies and can make traditional assets such as stock footage outdated. Text-to-video AI has the potential to revolutionize various creative sectors, including filmmaking, advertising, graphic design, and game development, alongside industries like social media, influencer marketing, and educational technology (Lohchab H. & Rekhi D., 2024).

In the domain of text-to-video artificial intelligence, numerous challenges remain unresolved. Among these, intellectual property protection is paramount, and systems must maintain transparency regarding the resources employed in model training. However, companies like OpenAI are not “open” at all and tend not to disclose their resource list. OpenAI Sora is probably trained with video content from stock video repositories and content creators for these sites will not benefit from the possible earnings from this product.

An interesting research question in the uncanny valley of GANs is "Based on what the perception of trustworthiness for the contents of reality synthetically generated by the GANs vary?" (Carnevale, Delgado & Bisconti, 2023). Synthetically generated content can be used for political or some other misuse. Trustworthiness has numerous criteria; sincerity, objectivity, source credibility, vividness, perceived salience, information believability, and overall persuasiveness. Additionally, it is essential to investigate the psychological
dimensions of human-AI interactions within these systems (Carnevale, Delgado & Bisconti, 2023). We need to devise blockchain solutions or similar technologies to achieve trust and traceability.

There is a concern regarding the potential misuse of such systems for generating deepfake videos and inappropriate content. Some of these can be deceptive, harmful, or offensive material. Meticulous configuration is essential to filter these, especially political or sexual content to safeguard human rights. Content filtering will help in producing videos according to ethical and legal standards. Moreover, these should be watermarked to facilitate the tracing of their origin. OpenAI and many companies are using Content Provenance and Authenticity (C2PA - https://c2pa.org/) standards to watermark the synthetically generated images. However, C2PA was not used with videos at the time of the writing.

The computational and environmental costs associated with training the model, as well as the computing expenses incurred by executing any text-to-video command, remain undisclosed and warrant further investigation. It is imperative to integrate energy efficiency measures across the entire lifecycle of AI systems, spanning from their inception in design and development to their deployment and ongoing maintenance. This can be achieved through the adoption of refined algorithms, engineered to optimize computational requirements, and the utilization of energy-conscious hardware architectures, such as low-power processors and specialized AI accelerators. Additionally, techniques like model pruning, quantization, and distillation offer avenues to mitigate the computational complexity of AI models, thereby reducing energy consumption. Dynamic resource allocation strategies, aligning computational resources with demand fluctuations, further contribute to energy conservation efforts. Leveraging edge computing infrastructure facilitates the execution of AI tasks in proximity to data sources, diminishing energy expenditure associated with data transmission and centralized processing. Moreover, enhancements in data center efficiency, encompassing improvements in cooling systems, power management practices, and integration of renewable energy sources, hold promise for reducing overall energy consumption. It is imperative to broaden the discourse beyond energy footprints to encompass water footprints as well, underscoring the importance of developing environmentally sustainable solutions. Accordingly, data centers should
formulate comprehensive strategies aimed at augmenting water efficiency, embracing sustainable water sourcing practices, instituting community-based water reuse initiatives, and actively engaging in water replenishment endeavours to foster a net-positive impact on water resources.

The technical report released by OpenAI characterizes Sora as a "video generation model serving as a world simulator." The report suggests that expanding this video generation model could potentially lead to the development of a comprehensive simulator capable of representing the physical world. OpenAI asserts that Sora's advancements may serve as a foundational step toward models that possess the ability to comprehend and simulate the real world, marking a significant milestone in the advancement of Artificial General Intelligence (AGI). It appears that further advancements are on the horizon in this domain, particularly with the anticipated releases of OpenAI's ChatGPT5 and Google's Gemini LLM products. These developments are poised to increase the integration of artificial intelligence into various aspects of our daily lives, thereby altering our approaches to jobs and our comprehension of concepts.

8. CONCLUSION

Sora's ability to generate 60-second videos, surpassing its competitors who are limited to creating short clips lasting only a couple of seconds, has the potential to establish new standards in the field. Additionally, the quality of the videos and the utilization of data-driven physics engines have enabled the creation of more realistic videos, particularly in the impressive representation of human anatomy. Sora represents a significant advancement in Artificial General Intelligence (AGI), with the potential to increase the integration of artificial intelligence into various aspects of our daily lives, thereby reshaping our approach to work and our understanding of concepts.

There is also significant competition in text-to-video products, and we will likely see improvements across all products. However, several notable challenges require attention, particularly the protection of intellectual property rights and ensuring transparency in model training resources. Measures are needed to prevent the creation of deep fake videos and inappropriate content, with a primary focus on safeguarding human rights. Reliable
watermarking on generated videos is necessary to trace their origins. Blockchain technology can enhance trust and traceability within this domain. Furthermore, generative AI processes should be optimized to reduce computational and environmental costs to ensure sustainability. Addressing these challenges requires interdisciplinary cooperation to develop robust frameworks, regulatory measures, and technological advancements aimed at promoting the ethical and responsible application of text-to-video AI.

The undisclosed computational and environmental expenses linked to both model training and the execution of text-to-video commands demand thorough scrutiny and exploration. The environmental consequences stemming from excessive computational activity underscore the need for meticulous examination. It is crucial to incorporate measures promoting energy efficiency throughout the entirety of AI system development, encompassing stages ranging from initial design and development to subsequent deployment and continuous maintenance. Furthermore, expanding the discourse to encompass not only energy footprints but also water footprints emphasizes the significance of cultivating environmentally sustainable approaches.
ACKNOWLEDGEMENT

The study is a part of LLM joint research between MSKU Metaverse Lab and Manisa Celal Bayar University.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHORS’ CONTRIBUTIONS

All authors have participated in drafting the manuscript. All authors contributed equally to the manuscript and read and approved the final version of the manuscript.

CONFLICT OF INTEREST

The authors certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript.

REFERENCES


Hilado, J. A. C. (2023). *Comparative Study of Text-To-Image Models: A Focus on Subject-Specific Training for Improved Generation* [Thesis, California State University]. https://scholarworks.calstate.edu/downloads/ns064f44h


Lumiere (2024) https://lumiere-video.github.io/


Sora (2024) OpenAI Sora. https://openai.com/sora


