Carbon in Motion: Characterizing Open-Sora on the Sustainability of Generative AI for Video Generation

Baolin Li
Northeastern University

Yankai Jiang
Northeastern University

Devesh Tiwari
Northeastern University

ABSTRACT

The rapid rise of generative AI (GenAI) technologies has brought innovative video generation models like OpenAI's Sora to the forefront, but these advancements come with significant sustainability challenges due to their high carbon footprint. This paper presents a carbon-centric case study on video generation, providing the first systematic investigation into the environmental impact of this technology. By analyzing Open-Sora, an open-source text-to-video model inspired by OpenAI Sora, we identify the iterative diffusion denoising process as the primary source of carbon emissions. Our findings reveal that video generation applications are significantly more carbon-demanding than text-based GenAI models and that their carbon footprint is largely dictated by denoising step number, video resolution, and duration. To promote sustainability, we propose integrating carbon-aware credit systems and encouraging offline generation during high carbon intensity periods, offering a foundation for environmentally friendly practices in GenAI.

KEYWORDS

Sustainable Computing; Generative AI; Video Generation.

1 INTRODUCTION

Generative AI (GenAI) is experiencing a significant surge in popularity. Following the launch of ChatGPT by OpenAI in November 2022, which amassed one million users within just five days, large language models (LLMs) that generate text responses to user prompts have captivated major information technology companies. This has sparked a trend of frequently released new LLMs, including Meta Llama, Google Gemini, Anthropic Claude, Snowflake Arctic, and IBM Granite. While LLM capabilities have grown tremendously over the past year, OpenAI, as a leader in GenAI, announced a revolutionary video generation model, Sora, in February 2024. This model can transform user text prompts into highly realistic videos [32]. As humans interact significantly with visual data alongside natural language, the launch of OpenAI Sora is expected to replicate the "ChatGPT moment" and ignite a wave of investment in video-focused multimodal generative model development across the technology industry.

As GenAI develops, the carbon emissions incurred in training and deploying such models require urgent intervention, as these workloads execute on integrated circuits that incur carbon emissions both during manufacturing and operation when powered by the grid. Notably, the energy consumption of global data centers is projected to reach 1,000 TWh by 2026 [16] due to AI growth, and the corresponding carbon emissions could account for 8% of global emissions within a decade [12]. In this work, we are particularly interested in the inference process, as inference is expected to dominate the AI computing cycles [7, 27, 42]. Previous works have investigated the carbon footprint of generative language models [7, 11, 26, 29], but they have distinctive architectures compared to text-to-video models, and there is a lack of systematic effort in addressing the sustainability challenges in Sora-like video generation applications.

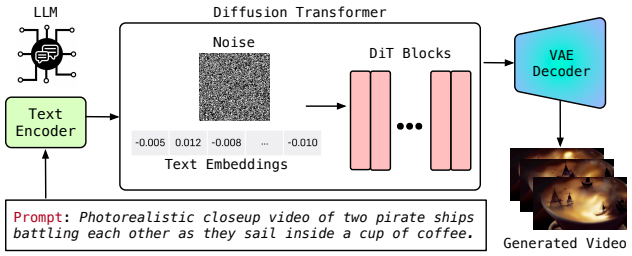

Figure 1: The process of video generation.

As video generation is poised to become the next milestone application of GenAI, we conduct a carbon-centric case study on this emerging field. **The contributions of this work can be summarized as follows:**

This work is the first to investigate the carbon footprint of video generation applications. While text generation applications use an autoregressive approach to generate text tokens, video generation applications rely on diffusion models that iteratively denoise a latent space video representation. By taking this first step, this work aims to shed light on the sustainability challenges of video generation.

Our characterization provides operational insights for making video generation services eco-friendly. Notably, video generation applications are significantly more carbon-intensive than text generation, with the primary source of emissions stemming from iterative diffusion denoising. We examine the carbon footprint and generation quality under various configurations of denoise step number, resolution, and duration.

We offer insights from this study for video generation service providers to integrate sustainability into their pricing models. Service providers should establish a carbon-aware credit system to incentivize environmentally friendly video generation practices. Furthermore, encouraging users to opt for offline generation during periods of high carbon intensity can significantly reduce emissions.

Next, we introduce the diffusion transformer-based video generation models and present our characterization details.

2 BACKGROUND AND METHODOLOGY

Video Generation Model Architecture. With the increasing popularity of transformers [38] and latent diffusion [36], video generation has widely adopted the diffusion transformer (DiT) [33] architecture, which is more scalable than the traditional U-Net architecture. The text-to-video generation process is illustrated in Fig. 1. Here, user prompts are tokenized and encoded using a

language model to create a noisy latent space representation. This representation is then fed into the diffusion transformer, which iteratively denoises it over a specified number of steps. Finally, a variational autoencoder (VAE) decodes the latent representation into video frames.

Carbon Footprint of Video Generation. The carbon footprint measures the greenhouse gas emissions, primarily CO_2, associated with the production and operation of computer components. For online applications, the carbon footprint comprises both embodied carbon and operational carbon [13]. Embodied carbon refers to the one-time emissions from manufacturing and packaging integrated circuit components. When calculating the carbon footprint for each video generation, the embodied carbon is proportionally allocated by dividing the generation time by the overall lifespan of the device. Operational carbon is calculated by multiplying the carbon intensity (gCO2/kWh) of the electricity grid by the energy used in the datacenter to power the inference server. The total carbon emissions associated with generating a video can be expressed as follows:

$$Carbon = Energy \cdot Intensity + \frac{T_{\text{gen}}}{T_{\text{life}}} \cdot Embodied \quad (1)$$

where T_{gen} and T_{life} represent the generation time and device lifespan, respectively. Note that this formulation assumes the hardware is busy serving requests throughout its lifetime, which may not always be true. However, one can adjust T_{life} accordingly and such formulation was used in existing application [18]. We model the embodied carbon of hardware devices using ACT [12] and modify the CarbonTracker [6] package to monitor the operational carbon.

Experiment Setup. Since OpenAI Sora is proprietary software, several open-source projects have attempted to replicate its video generation capabilities. Among them, we selected the Colossal-AI Open-Sora [43] model due to its popularity and its training similarity to Sora's description. Other projects, such as Open-Sora-Plan [23], are slower and currently lack multi-resolution/duration generation support. We established the inference benchmark using the latest Open-Sora v1.1.0 release on an NVIDIA A100 GPU (CUDA 12.1). To achieve optimal efficiency, we enabled FlashAttention [9] and xFormers [24] for acceleration. For our video generation benchmark, we used all the prompts from the OpenAI Sora demo [32] and Open-Sora examples.

3 CARBON FOOTPRINT ANALYSIS

In this section, we analyze the carbon footprint of video generation applications to address a series of unexplored research questions (denoted as RQs). These questions aim to uncover the environmental challenges associated with deploying video generation applications. By empirically studying them, we derive insights and implications for operating video generation services in an environmentally sustainable manner.

> **RQ 1.** *How does embodied and operational carbon account for the inference carbon footprint for video generation?*

This research question helps us understand the contributions of the manufacturing/packaging components and the inference server operation phase to the overall carbon footprint of the system. As outlined in Sec. 2, the embodied carbon of video generation depends solely on the time the request runs on the hardware as a fraction of the device's lifetime, while the operational carbon is determined by the energy consumption during generation and the carbon intensity of the grid. In this study, we assume a device lifetime of 5 years when calculating embodied carbon and a datacenter power usage efficiency (PUE) of 1.2 to calculate operational carbon, a typical value for efficiently operated datacenters [31]. All carbon numbers presented in this paper include both embodied and operational carbon.

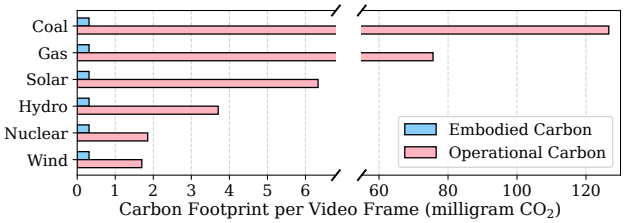

Figure 2: Carbon footprint when powered by the most commonly used energy sources in the US [3].

In Fig. 2, we quantify the carbon footprint of video generation when running the service using various energy sources that collectively account for over 97% of the electricity generated in the US [3]. The carbon intensity of each energy source is derived from ACT [12]. Our findings reveal that embodied carbon contributes negligibly to the overall carbon footprint compared to operational carbon. Even when the datacenter is powered by wind – the energy source with the lowest carbon intensity – operational carbon is 5.4 times the embodied carbon, accounting for 84% of the overall carbon footprint. When the system is powered by gas, the most common fuel source for electricity in the US, embodied carbon contributes only 0.4% to the carbon footprint of generating a video.

> **Insights and implications.** *The carbon footprint of video generation applications is predominated by its operational carbon, which depends on the local grid's carbon intensity. Therefore, as the inference service provider, it is a sustainable practice to deploy the application in datacenters powered by more renewable energy sources such as wind, nuclear, hydro, and solar. In addition, it is worth deploying the video generation models on hardware that costs more carbon to manufacture but has better power efficiency (e.g., chips with more advanced lithography).*

We have confirmed that the carbon footprint of video generation is dictated by its operational carbon. However, understanding the relative carbon emissions incurred during an inference request will help us gauge how carbon-demanding video generation applications truly are. This topic is discussed next.

> **RQ 2.** *How does the carbon footprint of video generation compare against text generation with a large language model (LLM)?*

Large Language Models (LLMs) have gained significant popularity, particularly in generative AI applications where the language model iteratively generates output tokens based on an input

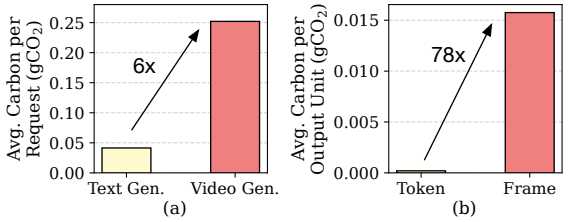

Figure 3: Video generation is significantly more carbon-demanding than text generation on LLMs. The shown example generates 2-second videos at 240p resolution.

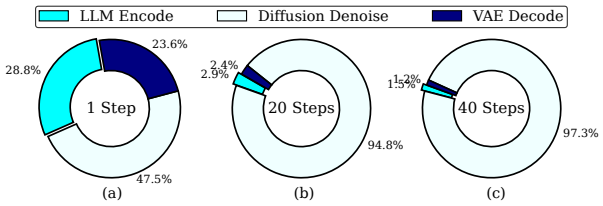

Figure 4: Carbon footprint of different phases of video generation when varying the number of diffusion denoising steps.

prompt. Here, we compare the inference carbon footprint of the Open-Sora model against the Meta Llama2 13B model. We chose an intermediate-sized Llama model with a memory footprint similar to that of the Open-Sora model. Note that we did not use Llama3 as its 13B variant had not been released at the time of this work. We understand this comparison can be dictated by the choice of model and configurations, but we try to make these numbers more representative by selecting popular workloads.

To evaluate the carbon footprint of the LLM, we used a mixture of representative language modeling datasets, including Alpaca [37], GSM8K [8], MMLU [14], Natural Questions [22], ScienceQA [28], and TriviaQA [17]. For video generation's carbon footprint, we generated 2-second videos at 240p resolution (we discuss the generation of longer videos at higher resolutions later in Sec. 3). A carbon intensity of 100 gCO_2/kWh is used in both cases. The goal is to compare the effort required to generate a relatively short, low-resolution video versus a text response to a prompt.

In Fig. 3 (a), we show the average carbon footprint per request for language generation compared to video generation. The average carbon used to generate a video is approximately 6× higher than that required to generate a sequence of text tokens. For a more comprehensive comparison of video generation versus text generation, Fig. 3 (b) provides a finer-grained analysis. Since videos are composed of frames, we quantify the carbon emissions corresponding to each frame. For text generation, responses are composed of tokens before being decoded into text, so we quantify the carbon emissions for generating each token (total inference carbon divided by the number of tokens generated, averaged across all dataset prompts). Fig. 3 (b) shows that the carbon per frame is about 78× higher than the carbon per token, highlighting that video generation is much more carbon-intensive than text generation. While extensive research has been conducted on the carbon efficiency of LLMs [7, 11, 26], our characterization emphasizes a more urgent need for sustainable practices in video generation.

Insights and implications. *Generating videos consumes significantly more carbon than generating text: the average carbon emission for a single 240p video frame is equivalent to generating 78 text tokens with comparably sized video and text generation models. As GenAI technologies advance towards multi-modality, we expect video generation to become a major contributor to the carbon footprint of GenAI in the future. Therefore, we call for more research focused on carbon awareness and sustainability in video generation applications.*

We have confirmed the carbon-demanding nature of video generation applications, but to identify the carbon bottlenecks, we need a deeper understanding of the video generation process. As introduced in Sec. 2, current state-of-the-art video generation applications use diffusion transformers that follow a general process: (i) using a language model to encode the user prompt into a latent space representation; (ii) gradually reversing the noise-adding process (learned from training) through iterative denoising; and (iii) converting the denoised latent space representation into video via the decoder of a variational autoencoder (VAE). Next, we examine the carbon impact of each of these processes during video generation.

RQ 3. *How do the text encoding, diffusion denoising, and video decoding phases account for the video generation carbon?*

In our video generation benchmark, following the architecture in Fig. 1, we use Google's T5 v1.1 xxlarge model, which has approximately 11 billion parameters, to encode the text prompt [35]. Note that this LLM is used solely for text encoding, a different process from text generation as discussed in RQ 2. For denoising, we utilize Open-Sora's spatial-temporal diffusion transformer (STDiT) model, allowing the diffusion model to iteratively refine its understanding of the input data, gradually reducing noise and improving signal fidelity. The number of denoise steps/iterations (typically tens to hundreds) can be adjusted during inference. The decoder is a variational autoencoder with KL loss [20] from the Huggingface Diffusers library. In this experiment, we continue generating 2-second videos at 240p resolution.

In Fig. 4, we first examine how much carbon footprint one diffusion denoising step has compared to the LLM encoding and VAE decoding phases. In Fig. 4 (a), we show that the carbon footprint of a single diffusion denoising step is comparable to the combined carbon footprint of the encoding and decoding phases. However, since the model requires multiple denoising steps, we increase the number of steps to 20 in Fig. 4 (b) and 40 in Fig. 4 (c). As the number of steps increases, it becomes evident that the diffusion denoising phase dominates the carbon footprint of video generation. Note that 40 denoising steps are considered small, as the Open-Sora model defaults to using 100 steps. Yet, even with just 40 denoising steps, 97.3% of the carbon footprint of generating short, low-resolution videos (2s, 240p) is already accounted for, rendering the encoding and decoding carbon negligible. It's worth noting that, even though we encode the user prompt using an LLM, this process only consumes about 60% of the carbon footprint of a single denoising step. This is because the denoising process also heavily utilizes attention operators that are widely used in language models. However, rather

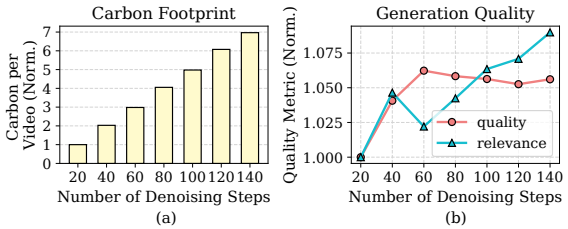

Figure 5: Carbon and video quality impact on adjusting the number of diffusion denoising steps.

than applying attention to text tokens, the attention operators are applied to the spatial and temporal features of the video.

> **Insights and implications.** *The carbon emissions of video generation applications are dominated by the iterative diffusion denoising phase, even when denoising is done in the latent space. The length of the input prompt has almost no impact on video generation carbon emissions, as the LLM text encoding carbon footprint is negligible. Given that video generation has a much larger carbon footprint than text generation, it would be a carbon-friendly practice for service providers to refine user prompts using language models before feeding them into the video generation model. For example, OpenAI could pass user prompts to GPT for text refinement, then feed them into their Sora model for higher-quality generation with minimal additional carbon emissions.*

Now that we have identified the carbon bottleneck in diffusion denoising, the number of denoising steps becomes a primary carbon factor. However, as discussed in previous work, the number of denoising steps (default is 100 in Open-Sora) is not a straightforward parameter to configure due to its complex interaction with video quality. Previous research has shown that a higher number of steps generally leads to better image quality during image generation [41]. For the next research question, we investigate the impact of the number of steps on Open-Sora's carbon footprint and video quality.

> **RQ 4.** *How should we control the number of diffusion denoising steps to account for both video quality and carbon emission?*

Measuring carbon emissions can be done by evaluating the quantities in Eq. 1. To assess video generation quality, we consider two separate perspectives. First, we evaluate video quality without any context – reflecting how realistic the video appears based on common sense. For instance, whether subjects flicker in the video. Second, we assess how well the video aligns with the provided text context. For instance, whether the video shows cats when users request dogs.

To quantify these two properties, we modify VBench [15], a state-of-the-art video quality benchmark collection. We select MUSIQ [19] to evaluate frame distortion in the video as a video quality proxy, and ViCLIP [40], a video extension of the OpenAI CLIP score [34], to measure the correlation between the generated video and the original prompt. A higher score indicates higher generation quality for both metrics. We acknowledge that these metrics can only serve as proxies for certain aspects of the video, as judging video quality is inherently complex and subjective.

Table 1: Latent space dimensions of Open-Sora video generation at various resolutions.

Resolution	Aspect Ratio	Height × Width	Latent Dimensions
144p	0.56	144 × 256	# of frames × 18 × 32
240p	0.56	240 × 426	# of frames × 30 × 53
360p	0.56	360 × 640	# of frames × 45 × 80
480p	0.56	480 × 854	# of frames × 60 × 106
720p	0.56	720 × 1280	# of frames × 90 × 160

In Fig. 5 (a), we show that the carbon emissions per video generation increase proportionally with the number of denoising steps. This is expected, as we demonstrated in RQ 3 that denoising dominates the inference carbon footprint and each denoising step repeats the same operations. However, when we examine video generation quality in Fig. 5, a different trend emerges. Video quality rises steeply when the number of steps is initially increased to 60, then begins to plateau, and even slightly declines as the number of steps continues to increase. The video's relevance to the input prompt, on the other hand, shows an upward trend as the number of denoising steps increases. In summary, increasing the number of denoising steps enhances video quality up to a certain point before hitting a plateau, while video relevance continues to improve. However, this also results in a proportional increase in carbon emissions.

> **Insights and implications.** *Service providers should standardize video generation quality metrics to help determine the appropriate number of denoising steps, balancing quality and carbon emissions. Since video generation carbon is primarily operational (RQ 1), the denoising procedure should be configured differently based on varying carbon intensity periods. We encourage diffusion model researchers to design carbon-aware denoising schedulers to promote sustainable practices in video generation applications.*

In real-world scenarios, videos would need to have much higher resolutions. We did not use higher-resolution examples to answer previous RQs for ease of visualization; otherwise, the carbon footprint of text generation and encoding/decoding would become invisible. Next, we investigate the scenario of video generation at varying resolutions and durations.

> **RQ 5.** *How do video resolution and length affect the carbon footprint of video generation?*

The Colossal-AI Open-Sora model is trained with videos of varying resolutions and durations. When generating videos at different resolutions and durations (number of frames), the latent space dimension changes accordingly, as shown in Table 1. Based on our previous discussions, the carbon footprint of video generation is dominated by the diffusion denoising process in the latent space. Therefore, changes in the latent representation dimension significantly impact the carbon footprint of video generation.

In Fig. 6, we show the carbon footprint per video generation at various video resolutions, normalized to 240p videos. As video height and width (pixels) scale, the dimensions of the latent space

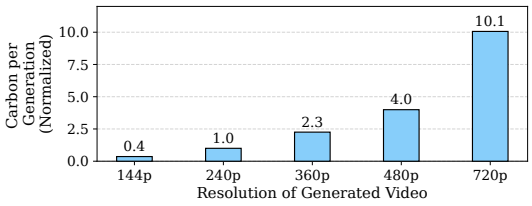

Figure 6: The carbon footprint of video generation scales almost quadratically with video resolution.

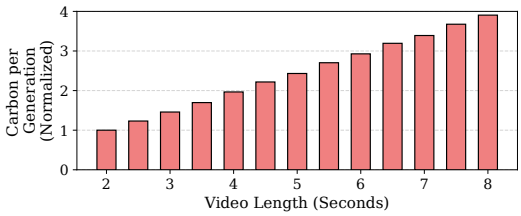

Figure 7: The carbon footprint scales linearly with generated video duration.

representation also scale linearly with height and width, respectively (Table 1), resulting in a quadratic increase in tensor size because two dimensions scale linearly. The carbon emissions in Fig. 6 follow a near-quadratic trend as video resolution scales, indicating that generating videos at higher resolutions incurs significantly higher carbon emissions. Notably, generating videos at 720p produces 10× more carbon than at 240p.

In Fig. 7, we fix the resolution and adjust the video generation duration from 2 seconds to 8 seconds, which is equivalent to increasing from 16 frames to 64 frames at a frame rate of 8 frames per second. As shown in Table 1, the latent space does not downscale the number of frames in the representation. Similarly, as observed in Fig. 7, the carbon footprint of video generation scales linearly with video duration.

> **Insights and implications.** *Users' generation requirements (i.e., resolution and duration) heavily impact the carbon footprint, especially when higher-resolution videos are requested. Service providers should guide users to conduct trials at lower resolutions before generating videos at high resolutions. Additionally, it's a carbon-friendly practice to forward high-resolution requests to datacenter regions with lower carbon intensity and de-prioritize high-resolution and long video requests during periods of high carbon intensity.*

4 DISCUSSIONS

In Sec. 3, we quantified the carbon impact of video generation applications using the latest Open-Sora model. Based on this characterization and the insights gained, we offer some takeaways and suggestions for video generation service providers to incorporate more carbon-friendly practices.

Adaptive Generation Credit System. Service providers often grant users a specific amount of generation credit and charge for each generation (e.g., Adobe Firefly [2]). We suggest that service providers implement a carbon-aware credit model to charge users based on their carbon usage. Specifically: (i) Since generation carbon is primarily operational (RQ 1), the provider should offer users generation discounts during periods of low carbon intensity. (ii) Users can specify their desired video resolution and duration, while the system should scale the generation credits accordingly to encourage carbon savings. Based on our characterizations in RQ 5, providers should scale credits quadratically when users request non-default video resolutions and linearly when they request non-default video durations. (iii) Providers should conduct A/B tests to study users' preferences for different numbers of denoising steps during generation and grant more credits to users who prefer a lower number of denoising steps.

Online to Offline Generation. Video generation applications typically require significantly more processing time for an online inference request compared to text or image generation applications. For instance, using the Open-Sora model to generate an 8-second video at 480p resolution takes approximately 8 minutes on an NVIDIA A100 Tensor Core GPU. Consequently, users generally do not expect video generation to respond as quickly as other generative AI applications, such as chatbots.

Recognizing this difference from other online services, we recommend that video generation service providers encourage users to convert their online processing requests to offline mode when the grid's carbon intensity is high. This approach is intuitive because users are unlikely to wait for extended periods for video generation, especially when multiple video samples are created from a single prompt for the user to select from. If users opt in, the provider can defer generation until the carbon intensity decreases or until the offline processing deadline, rewarding the user with discounts or generation credits. This is similar to how Amazon encourages customers to opt for longer delivery times to save trips and packaging [4].

5 RELATED WORK

Multiple studies have established carbon estimation models and design insights for reducing the carbon footprint [1, 10, 12]. Furthermore, carbon-aware design has extended to various computer science areas, including machine learning [27, 30, 42], cloud and supercomputing datacenters [5, 25, 39], and operating systems [21]. This effort has recently extended to generative AI [7, 11, 26, 29]. For example, LLMCarbon [11] introduces an end-to-end carbon footprint projection for LLMs. However, current research lacks an understanding of the carbon impacts of video generation, while our study reveals that video frames are significantly more carbon-intensive to generate than text tokens. Our work serves as a pioneer in investigating the carbon emissions of video GenAI.

6 CONCLUSION

In this paper, we quantitatively analyze the carbon footprint of video generation applications using the state-of-the-art Open-Sora model. Our characterization reveals that video generation could become the major carbon emission source in GenAI and proposes several environmentally sustainable practices. We hope our insights will help ML practitioners design more carbon-efficient video generation systems.

REFERENCES

[1] Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Maeng, Udit Gupta, Manoj Chakkaravarthy, David Brooks, and Carole-Jean Wu. 2023. Carbon explorer: A holistic framework for designing carbon aware datacenters. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*. 118–132.

[2] Adobe. 2024. *Create in new ways with generative AI powered by Adobe Firefly.* https://helpx.adobe.com/firefly/using/generative-credits.html

[3] U.S. Environmental Protection Agency. 2024. *Electric Power Sector Basics.* https://www.epa.gov/power-sector/electric-power-sector-basics

[4] Amazon. 2024. *Amazon Day Delivery.* https://www.aboutamazon.com/news/operations/what-is-amazon-day-delivery

[5] Thomas Anderson, Adam Belay, Mosharaf Chowdhury, Asaf Cidon, and Irene Zhang. 2023. Treehouse: A case for carbon-aware datacenter software. *ACM SIGENERGY Energy Informatics Review* 3, 3 (2023), 64–70.

[6] Lasse F Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint arXiv:2007.03051* (2020).

[7] Andrew A Chien, Liuzixuan Lin, Hai Nguyen, Varsha Rao, Tristan Sharma, and Rajini Wijayawardana. 2023. Reducing the Carbon Impact of Generative AI Inference (today and in 2035). In *Proceedings of the 2nd Workshop on Sustainable Computer Systems*. 1–7.

[8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).

[9] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems* 35 (2022), 16344–16359.

[10] Mariam Elgamal, Doug Carmean, Elnaz Ansari, Okay Zed, Ramesh Peri, Srilatha Manne, Udit Gupta, Gu-Yeon Wei, David Brooks, Gage Hills, et al. 2023. Carbon-Efficient Design Optimization for Computing Systems. In *Proceedings of the 2nd Workshop on Sustainable Computer Systems*. 1–7.

[11] Ahmad Faiz, Sotaro Kaneda, Ruhan Wang, Rita Osi, Parteek Sharma, Fan Chen, and Lei Jiang. 2023. LLMCarbon: Modeling the end-to-end Carbon Footprint of Large Language Models. *arXiv preprint arXiv:2309.14393* (2023).

[12] Udit Gupta, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin S Lee, David Brooks, and Carole-Jean Wu. 2022. ACT: Designing sustainable computer systems with an architectural carbon modeling tool. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*. 784–799.

[13] Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. 2021. Chasing carbon: The elusive environmental footprint of computing. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 854–867.

[14] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).

[15] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2024. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[16] IEA. 2024. Electricity 2024, analysis and forecast to 2026. https://www.iea.org/reports/electricity-2024

[17] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551* (2017).

[18] Sudarsun Kannan and Ulrich Kremer. 2023. Towards Application Centric Carbon Emission Management. In *Proceedings of the 2nd Workshop on Sustainable Computer Systems*. 1–7.

[19] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5148–5157.

[20] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[21] Sven Köhler, Benedict Herzog, Henriette Hofmeier, Manuel Vögele, Lukas Wenzel, Andreas Polze, and Timo Hönig. 2023. Carbon-Aware Memory Placement. In *Proceedings of the 2nd Workshop on Sustainable Computer Systems*. 1–7.

[22] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.

[23] PKU-Yuan Lab and Tuzhan AI etc. 2024. *Open-Sora-Plan.* https://doi.org/10.5281/zenodo.10948109

[24] Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. 2022. xFormers: A modular and hackable Transformer modelling library. https://github.com/facebookresearch/xformers.

[25] Baolin Li, Rohan Basu Roy, Daniel Wang, Siddharth Samsi, Vijay Gadepally, and Devesh Tiwari. 2023. Toward Sustainable HPC: Carbon Footprint Estimation and Environmental Implications of HPC Systems. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–15.

[26] Baolin Li, Yankai Jiang, Vijay Gadepally, and Devesh Tiwari. 2024. Toward Sustainable GenAI using Generation Directives for Carbon-Friendly Large Language Model Inference. *arXiv preprint arXiv:2403.12900* (2024).

[27] Baolin Li, Siddharth Samsi, Vijay Gadepally, and Devesh Tiwari. 2023. Clover: Toward Sustainable AI with Carbon-Aware Machine Learning Inference Service. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–15.

[28] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems* 35 (2022), 2507–2521.

[29] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research* 24, 253 (2023), 1–15.

[30] Priyanka Mary Mammen, Noman Bashir, Ramachandra Rao Kolluri, Eun Kung Lee, and Prashant Shenoy. 2023. Cuff: A configurable uncertainty-driven forecasting framework for green ai clusters. In *Proceedings of the 14th ACM International Conference on Future Energy Systems*. 266–270.

[31] NREL. 2024. *High-Performance Computing Data Center Power Usage Effectiveness.* https://www.nrel.gov/computational-science/measuring-efficiency-pue.html

[32] OpenAI. 2024. Creating video from text – Sora is an AI model that can create realistic and imaginative scenes from text instructions. https://openai.com/index/sora/

[33] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4195–4205.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[35] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[37] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html* 3, 6 (2023), 7.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[39] Jaylen Wang, Udit Gupta, and Akshitha Sriraman. 2023. Peeling Back the Carbon Curtain: Carbon Optimization Challenges in Cloud Computing. In *Proceedings of the 2nd Workshop on Sustainable Computer Systems*. 1–7.

[40] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. 2023. InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation. In *The Twelfth International Conference on Learning Representations*.

[41] Daniel Watson, William Chan, Jonathan Ho, and Mohammad Norouzi. 2021. Learning fast samplers for diffusion models by differentiating through sample quality. In *International Conference on Learning Representations*.

[42] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems* 4 (2022), 795–813.

[43] Zangwei Zheng, Xiangyu Peng, and Yang You. 2024. *Open-Sora: Democratizing Efficient Video Production for All.* https://github.com/hpcaitech/Open-Sora