the brandtech group



GENERATIVE AI ENVIRONMENTAL IMPACT STUDY: **VIDEO GENERATION**

JULY 2026

Table of Contents

Foreword 3

Introduction 4

Professional vs. Recreational Video 5

Calculation Methodology 7

Results 9

Case Study 12

Guidance and Best Practices 13

Appendix 15
How Video Generation Works
How We Kept Our Own Study Lean

Contributors 17

Foreword

Rebecca Sykes — **Partner, The Brandtech Group:**

“Our first report measured the carbon cost of the AI we use every day in text and image generation. This study extends that to video, where the energy demands are far higher. The question now isn’t whether marketers adopt AI - it’s how. Yet only a tiny proportion of total content production is currently being carried out with or by AI, so its overall footprint remains relatively small. At the scale we’re helping brands reach, it won’t. That’s why measurement has to sit where the work is made - in the tool, the workflow, the choice of which model runs and why - so the guardrails for lightest-impact use are built in early, not bolted on once the volume scales. It’s the standard we hold ourselves to, and the one we’re asking the industry to meet.”

Agathe Gourhannic — **Global Sustainability Manager, fifty-five:**

“Initially, we believed that the interpretations from our first chapter—such as the output size and number of generations—would simply apply to videos.
However, we quickly realized that video generation is infinitely more complex:

Our challenge has therefore been to concretely define actionable recommendations adapted to the professional use of AI for marketing.”

Ronan Robe — **Co-founder and CPO, fruggr:**

“For 2 years now, we have been challenging our methodologies together with fifty-five to make sure there are no blind spots in each of our work. Consequently, we readily agreed to cross-check their research on the carbon footprint of AI in order to provide brands with the precise operational KPIs needed to scale Gen AI responsibly without compromising their ESG commitments.”

Introduction

The rapidly advancing capability for video generation using AI represents a profound shift for marketers and creatives. What once required physical sets, crews, travel, and logistics can now be generated through a series of prompts.

But as we are increasingly aware, this transformational technology comes at a cost. AI is associated with a high demand for critical materials and substantial energy use, leading to serious environmental consequences. Furthermore, the increased energy demand fuels competition for electricity and strategic resources, placing additional strain on infrastructure and resource-constrained regions.

At The Brandtech Group, we are actively driving the AI creative revolution. Consequently, it is our responsibility to understand and mitigate the environmental footprint of these new tools and capabilities, not just as an ethical obligation but a necessity for responsible innovation.

This report builds on our foundational work analyzing the carbon footprint of AI-generated text and images. By synthesizing learnings from existing research with the results of our own experiments, our goal is to provide practical, data-driven insights to our partners and the wider marketing community on how their creative decisions translate into energy use and carbon emissions.

It is important to note that this study underwent peer review and critical assessment by fruggr, a company specializing in environmental impact assessments of digital activities with open-source methodologies.

Professional vs. Recreational Video

To assess the carbon footprint of AI-generated video, we should examine the context in which the technology is deployed. Iterative creative development for professional productions and casual experimentation involves different technical demands and can affect the carbon footprint of these projects by a factor of 30.

Professional production requires a high degree of precision, photorealism, and strict adherence to brand guidelines to ensure outputs are fit for public consumption. To achieve a brand-safe output with high satisfaction of quality, creatives often need numerous iterations in a multi-step, multi-tool process. Ideation, high-quality input images, short sequence video generation, refinement, and end-to-end assembly—this can accumulate into up to a week of work for an experienced creative. Additionally, professional-grade outputs demand high-definition inputs and outputs, often in 4K.

For example, we built a 30-second commercial video for an advertiser that required:

- **Prompt preparation**: 500 text generations
- **First frames** to ensure the visuals meet the client's expectations: 4000 image generations
- **Video generation** with start/end frames and upscaling: 926 video generations

At the same time, professional teams can also bring efficiencies that can materially influence how these tools are used. With stronger prompt engineering skills and a better understanding of model capabilities, experienced AI creatives and production teams can manipulate outputs more effectively and reduce trial-and-error generations.

Conversely, if a video is being generated for recreational purposes, the technical and contextual constraints are far less strict. Users will typically use general-purpose models, simplified workflows, and require lower-quality outputs. While each generation may consume fewer

resources, a lack of user expertise often results in experimentation and high trial-and-error. Coupled with the extremely low barrier to entry, this pattern of casual, individual use can still result in meaningful cumulative environmental impact.

For example, a 30-second general purpose video can be built with less than 50 video generations, without storyboards, prompt refinement, or upscaling. That is also the problem. When the barrier to creating AI video is so low, volume becomes the story: millions of users generating video at a pace and scale no one is tracking. The carbon cost of a single casual video is negligible. The carbon cost of a culture of casual generation is not.

The environmental footprint of AI video is heavily dependent on the context in which it is created, shaped by the intent, quality requirements, and workflow design. Therefore, any measurement also requires evaluating the full context of use and critically assessing if the purpose justifies the environmental impact generated.

Calculation Methodology

We focused on the carbon measurement of inference (i.e., the act of generating a video) and set aside the training phase (i.e., the preparation of the model before deployment). We have limited data available on training, and this follows the same principles we covered in the first iteration of this study.

Our approach blends the best existing research with our own experiments to build a robust picture:

1. **We started with the best work already out there**. We analysed two pioneering studies—Hugging Face (July 2025) and Northeastern University (December 2024)—to build a strong foundation and pull our first reliable data points on video's carbon footprint.

2. **We ran our own tests to fill the gaps.** We chose two common, open-source models that stand in for the closed models marketers and creatives use: Stable Video Diffusion and CogVideo. We configured a virtual machine to generate video on these models and used codecarbon.io to track electricity use and CO2 emissions. We ran multiple tests to isolate the effects of individual variables (duration, number of characters, type of input, etc).

3. **We compared the full spectrum of data** from external studies and our own experiments.

4. **We built our calculator.** We combined the external and internal data points into a common dataset and built an algorithm on top of it. For models for which we had data, we extrapolated to account for resolution, denoising steps, country, duration, and FPS. For models for which we had no data, the model inherits data from others sharing its architecture (DiT, UNet, or Hybrid). Access the model list here.

IMPORTANT CAVEATS

Our study extrapolates carbon footprint based on limited experiments. We cannot guarantee exact results for every model or untested scenario. To review our data points and extrapolation methods, please refer to our calculator data source and GitHub repo.

To ensure complete transparency regarding our environmental metrics, our framework relies on the following methodological boundaries:

- **Water Consumption (WUE):** The measured water footprint focuses exclusively on direct server cooling water and does not include indirect water usage tied to the regional electricity production mix.
- **Hardware & Server Scope:** Our calculation approach is centered on the primary compute component (GPU) and does not constitute a full Life Cycle Assessment (LCA) of the host server.
- **GPU Utilization & Efficiency:** GPU utilization is estimated based on the overall load observed during our laboratory Virtual Machine (VM) tests, which may not reflect the multi-tenant optimization typical of large SaaS clusters.

It's also important to note that our tests were run on single machines. Model providers operating large data centers get economies of scale and efficiency gains that cannot be replicated or factored into our data. We also focused only on the measurement of data center carbon footprint, omitting network and end-device impact as we consider them negligible by comparison.

Results

GENERATING VIDEO IS AN INTENSIVE AND EXPENSIVE PROCESS.

Video generation is considerably more demanding than text or image generation. The evidence came early. Our tests produced nothing usable until we moved to more powerful data centers[1], indicating that video generation requires the most intensive machines.

The numbers continue to prove this. Generating a 2-second video uses three times the electricity of generating a 1,000-word text[2]—roughly 30 Wh per clip—on average across models tested. As with text and image, the carbon footprint of video is highly dependent on the carbon intensity of the machines behind it. For example, this same 2-second video clip could generate almost 12 gCO2e with servers in the US (the equivalent of 10 web searches) vs 1 gCO2e in France[3].

There is a cost dimension as well. We estimate that electricity alone runs at €4 in machine cost on a standalone setup. This price is higher than what model providers currently charge us, which is no more than €0.30 per 2-second clip. That gap between what generations cost vs what the consumer is charged is being absorbed by providers for now. As adoption grows, it is reasonable to expect them to close the gap and pass the cost on. A more frugal approach to AI is not only a sustainable position, but also a hedge against pricing that could undermine the economics of a business model built on inexpensive generations.

THESE VARIABLES DIRECTLY IMPACT CARBON FOOTPRINT.

Certain parameters have a measurable impact on a video's carbon footprint. Knowing which variables, and which direction they push, is where marketers can make a meaningful difference.

Sources:
1) Nvidia A100 GPUs (versus 2 x Tesla T4)
2) Text generation: 7 Wh for a 1000-word output on GPT-4o mini (ref Brandtech's Generative AI Environmental Impact Study)
3) Emissions factors here

Output Resolution: generating a higher-resolution video means more computation requirements—and the relationship is not linear but quadratic. According to Northeastern University's research, doubling the resolution can quadruple the energy used. The challenge is many consumer-facing models are configured with a fixed default resolution that users cannot change.

Duration and frame rate: the cumulative volume of frames generated—determined by the length of the clip and its frames-per-second (FPS)—is proportional to the computational workload and the resulting carbon emissions. As the frame count rises, the environmental impact scales accordingly. Most video generation tools operate at a standard default of 24 FPS.

Some models, like CogVideo, work differently as they always produce a fixed block of frames (for example, 48 frames from 6 seconds x 8 FPS) and then cut the video to match the user's request. In these cases, the carbon footprint remains the same even with a shorter duration request as the work has already been done.

Number of denoising steps: most people have never encountered this variable. "Steps" refers to the number of passes a model[4] takes to turn an initial field of visual noise into a finished frame. The effect resembles a snow globe that progressively becomes clearer and clearer.

As observed in our tests and by Northeastern University, the carbon footprint increases proportionally to the number of steps. Typically, this cannot be adjusted unless the user is running an open-source model and configuring it manually.

Text-to-video vs. image-to-video: the more efficient choice of input—text or a first image frame—is highly dependent on how the model was built.

Sources:
4) Diffusion models

For example, Stable Video Diffusion (SVD) is built on the image model Stable Diffusion, so it always starts by generating an image first. Providing SVD a first frame image as input is approximately 20% less intensive than a text input alone.

Conversely, CogVideo (CV) must encode any input images into video format, so starting with a first frame image input is approximately 40% more intensive than a text-only prompt. Google's Veo3 model generates in the same manner[5].

Despite these different treatments of first frames, using them can significantly reduce the number of iterations required to achieve the expected output. So it would be worth using a first frame to reduce the overall carbon footprint of the project even if the footprint of the individual generation is higher.

And several settings have no effect at all. The most striking efficiency finding is that each generation and each sample requires the same work, whether run alone or bundled with others. There is no computational or energy efficiency gain from generating multiple videos in a single prompt. This is why it is critical to prepare carefully and refine the prompt before the first generation to minimize the need for multiple runs.

There are also a number of parameters that have no measurable impact on carbon footprint but can affect output quality:

- Input resolution
- Prompt format (json or natural language)
- Negative prompt
- Character(s)
- Visual style
- Guidance scale

The conclusion: a video's footprint comes down to resolution, frames, and steps — and, critically, how each individual model handles them.

Sources:
5) The Anatomy Of Veo3: Deepmind's Audiovisual Diffusion Model (July 9, 2025)

Case Study

Consider the 30-second commercial for our client referenced earlier. This required 500 text, 4,000 image, and 926 video generations, which totaled 120 kgCO2e. This is roughly the carbon cost of manufacturing one laptop[6].

The distribution of generations is notable here. Video generations accounted for under 20% of the total work, but contributed to 93% of the emissions.

30-Second AI-Generated Commercial | Emissions Breakdown				
Task type	**Generations**	**% of Generations**	**kg CO2e**	**% of CO2e**
Text	500	9%	1.601	1%
Image	4,000	74%	6.640	5%
Video	926	17%	112.803	93%
Total	**5,426**	**100%**	**121.044**	**100%**

This asymmetry comes down to how a video is actually generated. A text or image generation completes in a single inference pass. Video generation instead produces a sequence of frames over a rendering process that runs much longer, multiplying the compute, and therefore the emissions.

These emissions compound quickly at scale. A brand producing ten similar commercials over a year would generate over a ton of CO2e from generation work alone, equivalent to manufacturing roughly ten laptops. And this figure only accounts for the AI generation itself. In practice, most brands aren't replacing a commercial shoot with AI outright—they're combining the two, using AI to support pre-production, generate variants, or extend a set of live-action assets. The generation emissions sit on top of whatever traditional production is still involved, and a full accounting must include both.

Sources:
6) 156kgCO2e according to Base Empreinte ADEME (French Agency for Ecological Transition)

Guidance and Best Practices

Our findings point to a set of actionable recommendations for anyone producing AI videos—marketers, creatives, or anyone producing for personal use—to reduce the carbon footprint of their Gen AI video production. These recommendations echo our guidance for text and image generation (see Brandtech's Generative AI Environmental Impact Study), but video's added complexity means each lever requires adaptation.

01 **Use AI video only when it is warranted.** Given the footprint, it is necessary to question the final use of any AI generated videos. Does this video need to exist? Only generate when it is useful and well-considered, rather than by default.

02 **Select a low-carbon region.** As we saw earlier, the carbon footprint of AI generations is largely determined by the electricity source. Therefore, selecting a low-carbon region can significantly lower the carbon impact.

- Deploy an open-source model in a data center of your choosing in a low-carbon region
- Select a private model whose data centers are in a low-carbon region
- Choose a low-carbon region in your model's configuration, where that option exists

03 **Reduce the number of generations.** This is the lever most within your control. There are best practices you can put into place:

- Refine your prompt. You can consider using AI to optimize your prompt as a text generation. For example, if you are running 30+ iterations for a 6-second clip, your prompt needs refining. Rule of thumb: three unguided attempts > two attempts plus one AI-optimized prompt
- Keep "test and learn" to a minimum with video generation. It is the most carbon costly place to experiment
- If you are expecting a highly specific output, provide the model with a first frame image to reach the

expected result in fewer attempts. Note: for some models like Veo3, image generation + image encoding by the model + video generation can total more than generating from text alone, depending on how many attempts are required. In these cases, it may be more beneficial to focus on prompt refinement

04 **Draft in low-resolution.** Conduct all testing, storyboarding, and style refinement at the lowest output resolution available. Increase it only once the creative is approved and you are producing the final, production-ready output. Given the quadratic effect resolution has on carbon footprint, this can make a significant difference.

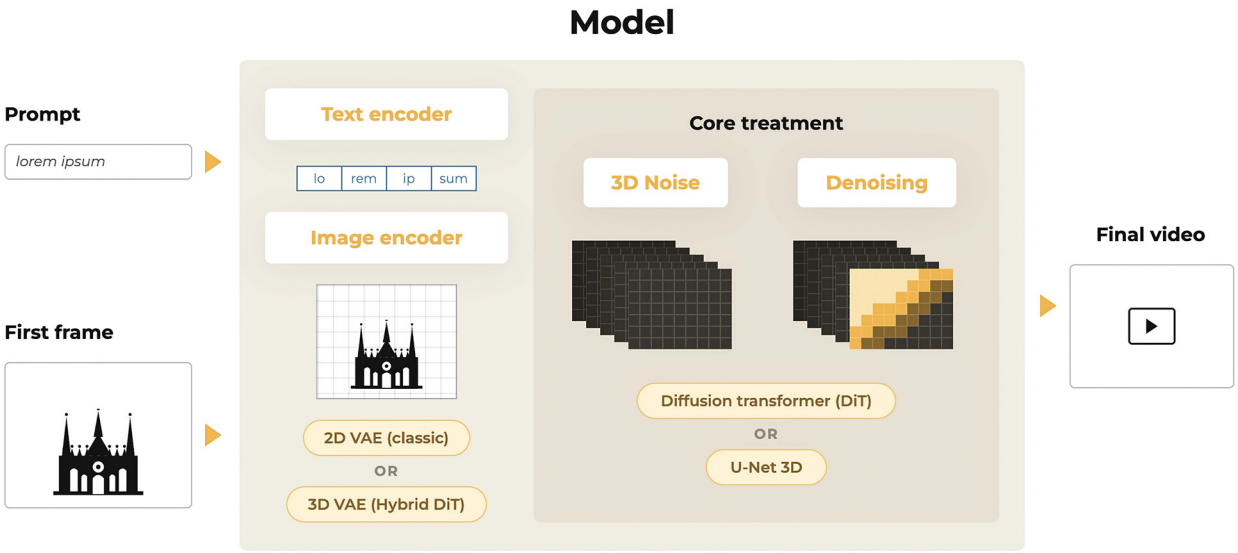
05 **Limit video duration.** Keep duration to the minimum, or to the model's default, and assemble several short clips rather than generating one long one. In practice, long videos are often wasted effort as only the first few seconds tend to be useful before artifacts or loss of coherence appear. If you generate at high volume, it is helpful to understand how your model behaves, since some models like CogVideo (unlike Veo3 or Stable Video Diffusion) always produce a long video before trimming it. In this case, stick to the model's default video duration.

06 **Limit denoising steps.** Favor models with fewer steps and, when customization is possible, keep the default settings to prevent unnecessary computation load. Counterintuitively, a higher number of steps does not guarantee better video quality. Both our tests and Northeastern University's study confirmed that quality tends to plateau well before the maximum number of steps is reached, so additional steps consume computation for no benefit.

Appendix

HOW VIDEO GENERATION WORKS

At a high level, every video generation model performs the same task—converting visual noise into a coherent, moving image through a process called denoising. This iterative process is what consumes the majority of the computational power and, consequently, drives the carbon footprint.



The architectural design of the model dictates how the denoising step is executed, which in turn shapes how it responds to the variables discussed above.

1. **The Denoising Step**: once the user's inputs (either text or an image) are encoded, the model iteratively predicts and removes noise. The method varies:
    - **Diffusion Transformer (DiT)**: used in models like Sora, these models refine all video token simultaneously at each step
    - **U-Net 3D**: used in models like Stable Video Diffusion, these models refine tokens progressively in batches

2. **The Image Encoding Step**: this step prepares input images for denoising. Some diffusion transformer models, like Veo3, use a more efficient 3D VAE (Variational AutoEncoder) to encode the entire video volume at once, rather than the traditional 2D VAE that processes each frame separately.

Beyond these shared structures, each model has its own unique history and "secret sauce" that makes it more or less sensitive to changes in input parameters. There is no universal rulebook.

HOW WE KEPT OUR OWN STUDY LEAN

To minimize the footprint of our own research, we implemented several eco-design choices:

- We tested only two models, drawing on existing research rather than replicating it
- We hosted experiments in low-carbon regions where possible
- We limited tests to short, 2-second videos to reduce run time

Note, we did have to upgrade our machine to larger GPUs to achieve satisfactory results.

Contributors

Agathe Gourhannic
Global Sustainability Manager, fifty-five

Ruga Amet Samatana
AI & Data Technical Lead, fifty-five

Urbain Dumas
Data Scientist, fifty-five

Ludovic Moulard
Independent Sustainability Expert

Hortense Juillet
Webanalytics and Sustainable AI Manager, fifty-five

Sharanya Sathyaprakash
AI Operations Manager, The Brandtech Group

James Dow
Gen AI Creative Director, The Brandtech Group

Jean-Pascal Martin, Ph.D.
Senior Consultant in AI and Responsible Digital Technology, fruggr