REFERENCE DOCUMENT

Estimating the Inference Carbon Intensity of AI Video Generation Models: Methodology, Estimates, and Justification

Prepared by: **InferenceCarbon.ai Reference Team**

Date: 25 March 2026 | Version: 1.8 | Classification: For publication — subject to stated caveats

Important note: *Video generation is the most energy-intensive and least-measured AI modality. The only direct measurements are for CogVideoX, an open-source model, measured by Luccioni/MIT Technology Review (2025). Commercial video models — Sora 2, Runway Gen-4/4.5, Kling, and Gemini Video (Veo) — have never been independently measured. The 31× energy difference between CogVideoX's low and high quality modes illustrates the enormous sensitivity to model architecture and quality settings. All commercial estimates in this document are engineering guesses with very low confidence.*

Scope notice: *This document estimates operational inference emissions only (Scope 2, electricity consumption during model inference). It does not constitute a lifecycle assessment and excludes: embodied emissions of hardware manufacturing; model training energy; upstream supply chain emissions (Scope 3); water consumption; and electronic waste. The estimates cover electricity consumed by GPU servers, supporting infrastructure (CPUs, memory, networking), and data centre overhead (cooling, power distribution), expressed via Power Usage Effectiveness (PUE).*

1. Executive Summary

This document provides carbon intensity estimates for AI video generation models, expressed in grams of CO_2 equivalent per second of generated video (gCO_2e/s). The primary anchors are the Luccioni/MIT Technology Review measurements of CogVideoX: 109,000 joules for 5 seconds of low-quality video and 3,400,000 joules for 5 seconds of high-quality video. Estimates are reported as Clean Energy Adjustment Factor (CEAF)-adjusted figures for decision support, with location-based gross figures for Scope 2 transparency.

Scaling methodology: Commercial video models cannot be directly measured, so we need to adopt a 'black box' approach estimated from publicly available information. This paper applies a hybrid two-approach methodology (Section 4.3). Approach A back-derives effective server power from the CogVideoX anchor (~950W on a single H100) and scales via API rendering time and an explicit GPU cluster size assumption. Approach B scales via resolution and frame rate ratios with sqrt() moderation. Where the two approaches diverge significantly — as they do for all commercial models, primarily due to unknown GPU cluster sizes — the geometric mean is used as the central figure. This is a material methodological limitation and all commercial estimates carry Low or Very Low confidence. The GPU cluster size assumption is the dominant source of uncertainty.

Model	Location-BasedRange(gCO_2e/s, Estimated)	Location-Based Central (gCO_2e/s, Estimated)	CEAF %	CEAF-Adjusted Central(gCO_2e/s)	Confidence
CogVideoX (High, 16fps)	45–110	69.9	0%	69.9	Medium
CogVideoX (Low, 8fps)	1.5–3.5	2.24	0%	2.24	Medium
Gemini Video (Veo)	3–180	104.8	66%	35.7	Very Low
Kling	9–470	141.7	0%	141.7	Very Low
Runway Gen-4/4.5	7–420	66.6	0%	66.6	Very Low
Sora 2	9–420	102.6	25%	76.9	Low
Sora 2 Draft	5–130	79.3	25%	59.5	Low

Table 1: Summary estimates. All figures per second of generated video at default quality. Location-based at per-provider grid intensity (see Section 3). CEAF: Google 66% (hourly CFE); OpenAI/Azure+Stargate 25% (blended: Azure annual match partially offset by Oracle Stargate's zero clean energy disclosure; see Section 3 and 7); CogVideoX, Kling, Runway 0% (open-source/undisclosed). Sorted alphabetically by model. Note: location-specific grid intensities are discussed in Section 4 as a future methodology improvement.

2. What Has Been Measured

Luccioni / MIT Technology Review (May 2025): Luccioni measured GPU energy for two versions of CogVideoX using CodeCarbon on NVIDIA H100 hardware. MIT applied a 2× overhead multiplier for total data centre energy (see MIT Technology Review methodology companion, 20 May 2025).[1][2]

CogVideoX Low Quality (8 frames per second (fps), ~5s): 109,000 J total. Per second: 21,800 J/s = 6.06 Wh/s.

CogVideoX High Quality (16fps, 5s): 3,400,000 J total. Per second: 680,000 J/s = 188.9 Wh/s.

The 31× difference reflects both higher frame rate (2×) and a substantially larger model (≈15× additional compute).

SINTEF (April 2024): An opinion piece (not a measurement). Estimated 60-second Sora video requires "1,800 image generations worth of compute." This is napkin math from Sora's pre-release technical paper, not an energy measurement.[3]

Chung & Chowdhury (2026): Found video generation sometimes consumes >100× the energy of images, and GPU utilisation differences can result in 3–5× energy differences between configurations.[4]

Unmeasured models: Sora 2, Kling, Gemini Video (Veo), and Runway Gen-4/4.5 have never been measured.

[1]MIT Technology Review, "We did the math on AI's energy footprint," 20 May 2025. CogVideoX Low (8fps): 109,000 J/5s; High (16fps): 3,400,000 J/5s. https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/
[2]MIT Technology Review, Methodology companion, 20 May 2025. Luccioni measured GPU via CodeCarbon; MIT 2× for data centre overhead. https://www.technologyreview.com/2025/05/20/1116331/ai-energy-demand-methodology/
[3]Eidnes, S., Meltzer, S. & Riemer-Sørensen, S. (2024) "What should we be using our electricity for?" SINTEF, 17 April 2024. Estimate: 60s Sora ≈ 1,800 image generations. https://www.sintef.no/en/latest-news/2024/what-should-we-be-using-our-electricity-for/
[4]Chung, J.-W. & Chowdhury, M. (2026) "Where Do the Joules Go?" arXiv:2601.22076. Video >100× image energy; GPU utilisation 3–5× differences. https://arxiv.org/abs/2601.22076

3. Provider Infrastructure and Grid Context

Provider	Model(s)	Infrastructure	Primary Data Centre Locations	CEAF %
Open-source	CogVideoX	User-deployed (H100)	Varies	0%
OpenAI	Sora 2, Sora 2 Draft	Azure + Oracle Stargate	US East, Texas (Abilene)	25%
Kuaishou	Kling	Chinese data centres	China	0%
Google	Gemini Video (Veo)	Google Cloud TPUs	US, EU, APAC	66%
Runway	Gen-4 / Gen-4.5	Undisclosed	Unknown	0%

Table 2: Provider infrastructure. OpenAI's multi-cloud shift (Azure + Oracle Stargate + AWS) means the blended CEAF of 25% reflects the Azure portion (50%) offset by Oracle Stargate (0%, no published clean energy data). Kling runs on Chinese data centres (Kuaishou) with no clean energy disclosure (CEAF Tier 3 = 0%). Runway's infrastructure is undisclosed (CEAF Tier 3 = 0%). Google CEAF from published 66% hourly CFE (Google 2025 Environmental Report). OpenAI/Azure from 100% annual renewable match (Microsoft, Feb 2026); blended to 25% to reflect Stargate fraction.

4. Anchor Derivation

Grid carbon intensity: Location-based estimates in this document use per-provider grid carbon intensity. The previous version used a uniform 370 gCO_2e/kWh (IEA's 2023 annual average for the United States electricity grid (IEA, Emissions Factors, 2024 edition). This is the same figure used across the InferenceCarbon paper series for consistency. The 370 gCO_2e/kWh figure is appropriate for models whose data centres are predominantly US-based (OpenAI/Azure/Stargate, Runway) when the precise regional grid and workload routing are unknown.[5]

Location-specific grid intensities: This version applies the grid intensity of each provider's actual data centre locations. EPA eGRID 2023 and Electricity Maps provide subregional data. Indicative values relevant to this paper: Virginia (RFCE, primary Azure US East) ~250 gCO_2e/kWh (nuclear-heavy); Texas/ERCOT (Oracle Stargate Abilene) ~380 gCO_2e/kWh (gas and wind); Iowa/Central US (Google MROW) ~350 gCO_2e/kWh; China national average (Kling) ~530 gCO_2e/kWh.[6]

Adopting location-specific grids has material effects: OpenAI's Azure (Virginia, ~250 gCO_2e/kWh) would see location-based estimates reduced ~32% vs the US average; Kling (China, ~530 gCO_2e/kWh) would increase ~43%. Applying location-specific intensities consistently across the full InferenceCarbon series is planned for a future methodology update. For this version, 370 gCO_2e/kWh is retained for cross-paper consistency, with location-specific figures noted in each model section where they are material.

4.1 CogVideoX Low Quality

Energy: 109,000 J ÷ 5s = 21,800 J/s = 6.06 Wh/s

Location-based (370 gCO_2e/kWh): 6.06 × 0.370 = **2.24 gCO_2e/s**

4.2 CogVideoX High Quality

Energy: 3,400,000 J ÷ 5s = 680,000 J/s = 188.9 Wh/s

[5]IEA, Emissions Factors 2024 (2024 edition). US electricity generation CO2 intensity for 2022: ~370 gCO2/kWh. Corroborated by Ember Global Electricity Review 2024, which reports US at 369 gCO2/kWh for 2023. https://www.iea.org/data-and-statistics/data-product/emissions-factors-2024; https://ember-energy.org/latest-insights/global-electricity-review-2024/

[6]EPA eGRID 2023 subregional emission rates. RFCE (Virginia): ~250 gCO2e/kWh; ERCT (Texas): ~380 gCO2e/kWh; MROW (Iowa/Central US): ~350 gCO2e/kWh. https://www.epa.gov/egrid. China national average from Ember Global Electricity Review 2024: 581 gCO2/kWh (direct); ~530 gCO2e/kWh (generation-weighted estimate used in this paper).

Location-based (370 gCO_2e/kWh): 188.9 × 0.370 = **69.9 gCO_2e/s**

4.3 Scaling Methodology

This section describes the complete scaling methodology. It draws on the approach used in the InferenceCarbon image model paper — back-deriving effective server power from a measured anchor, then applying energy = power × time — and extends it to address specific challenges in video generation.

Aligning with the image paper approach.

The image model paper derives an effective server power from its anchor (SD3 Medium on H100: 0.634 Wh in approximately 6 seconds → ~380W effective server power). Because image diffusion models run on a single GPU with near-constant utilisation throughout denoising, energy ≈ power × time applies cleanly, and API generation time becomes the primary scaling variable for all commercial models. The same physical principle holds for video: diffusion-based video models exhibit sustained GPU utilisation during denoising. Energy = server_power × rendering_time should therefore be valid.

Back-deriving server power from the CogVideoX anchor.

CogVideoX-5B (High Quality) is a 5B-parameter model that fits on a single H100 80GB. Community benchmarks on single-H100 hardware report generation times of approximately 30–90 minutes for a 720×480, 16fps, 5-second clip; central estimate: ~60 minutes (3,600 seconds).[7]

Implied effective server power = 3,400,000 J ÷ 3,600 s = 944W ≈ 950W.

This is physically consistent with: H100 GPU at ~350W average during sustained inference + server overhead (CPU, memory, networking) ~150W, with the 2× MIT data centre multiplier applied: (350 + 150) × 2 ≈ 1,000W. We adopt 950W as the implied effective server power for a single H100 server running video generation. Sensitivity range: 30–90 min rendering → 630W–1,900W effective server power.

The multi-GPU problem: why commercial video differs from image generation.

For image generation, single-GPU operation per inference request is the typical and reasonable assumption — which is why rendering time transfers cleanly to an energy estimate. Commercial video models almost certainly run on multi-GPU clusters for API inference. The number of GPUs per inference request is not disclosed for any commercial video model, and it dominates the energy calculation:

Example: Sora 2 generating a 10s clip in 120s wall-clock time.

At 8×H100 (×950W): 8 × 950W × 120s = 912,000 J = 25.3 Wh per clip = 2.5 Wh/s output.

At 64×H100 (×950W): 64 × 950W × 120s = 7,296,000 J = 202.7 Wh per clip = 20.3 Wh/s output.

An 8× difference in GPU count produces an 8× difference in energy from the same generation time. This is the fundamental limitation of applying the image paper's rendering-time approach directly to commercial video: without knowing the cluster size, rendering time cannot be converted to energy without this additional assumption.

The hybrid methodology.

Given this, we apply two approaches for each commercial model and compare them:

Approach A — Rendering-time-based (primary where GPU count is estimable). Energy/s_output = N_GPUs × 950W × hardware_correction × rendering_ratio ÷ 3,600, where rendering_ratio = rendering_seconds per output_second, and N_GPUs is an explicit stated assumption. This approach captures model size and architecture differences that Approach B cannot, but introduces GPU count as an additional major uncertainty.

Approach B — Resolution and frame rate scaling (cross-check). Energy/s_output = Anchor × sqrt(fps_ratio × res_ratio) × hardware_correction. This approach is architecture-agnostic and does

[7]CogVideoX-5B model card, Tsinghua University / THUDM. 5B-parameter video diffusion model; fits on single NVIDIA H100 80GB GPU. Community benchmarks report 30-90 min generation time for 720x480, 16fps, 5s clip. https://huggingface.co/THUDM/CogVideoX-5b

not require a GPU count assumption, but it does not capture differences in model size, number of diffusion steps, or inference cluster configuration. It is best understood as a pixel-compute proxy rather than a true energy predictor.

Where both approaches produce estimates within a factor of 2 of each other, we report the geometric mean as the central figure. Where they diverge more significantly, we report both and flag the methodological tension explicitly. The GPU count assumption is stated in Step 3 of each model section and is the dominant source of uncertainty for all commercial estimates.

The nine steps applied to each model.

Step 1: Reference anchor. CogVideoX High = 188.9 Wh/s output; implied server power ~950W on a single H100 (above).

Step 2: Rendering time. API generation time in seconds, converted to rendering-seconds per output-second. Non-compute overheads (queue, safety filtering) are typically under 5 seconds and are not discounted, making estimates slightly conservative.

Step 3: GPU cluster size assumption. Explicit stated N for the number of H100-class GPUs assumed to serve one API inference request. This is the primary uncertainty for all commercial video estimates.

Step 4: Energy from rendering time — Approach A. Energy/s_output = N × 950W × hardware_correction × rendering_ratio ÷ 3,600.

Step 5: Frame rate and resolution cross-check — Approach B. fps_ratio = fps_target ÷ 16. res_ratio = (W × H) ÷ (720 × 480). Quality factor = sqrt(fps_ratio × res_ratio). Cross-check Wh/s = 188.9 × quality_factor × hardware_correction.

Step 6: Hardware correction. H100/H200: ×1.0. Google TPU v5p/v6: ×0.25–0.6, central 0.5 (2–4× more efficient per FLOP). Unknown GPU-class hardware: ×1.0 (unknown direction of error).

Step 7: Central Wh/s estimate. Where Approach A and B are within a factor of 2: geometric mean. Where they diverge more: wider range with note on methodological tension.

Step 8: Location-based gCO_2e/s = Wh/s × provider CIF (per-provider grid intensity).

Step 9: CEAF-adjusted gCO_2e/s = location-based × (1 − CEAF%). See Section 7.

Multi-generation workflow note.

Some video workflows involve generating multiple draft clips before committing to a final full-quality render — analogous to Midjourney generating four candidates before upscaling the chosen one. Where a model explicitly supports a draft→final workflow (currently Sora 2 Draft → Sora 2), the total session carbon cost is:

Total gCO_2e = (N_drafts × Draft_gCO_2e/s × draft_duration) + (Final_gCO_2e/s × final_duration)

Example for Sora 2 (3 × 5s drafts then one 10s final, CEAF-adjusted at 25%): (3 × 62.9 × 5) + (81.3 × 10) = 944 + 813 = 1,757 gCO_2e ≈ 2.2× the final render alone. Per-second figures in the executive summary do not include a workflow multiplier; session-level accounting should apply it.

5. Per-Model Estimates

All commercial estimates are derived using the hybrid two-approach scaling methodology described in Section 4.3, anchored to CogVideoX High (188.9 Wh/s on a single H100). For each model, Approach A (rendering-time-based) and Approach B (resolution/frame-rate scaling) are computed independently; the geometric mean is used as the central figure where the two approaches are within a factor of 2. CogVideoX Low and High are directly measured. Ranges reflect the wider of ±50% uncertainty or the Approach A–B divergence for commercial models.

Per-provider grid intensity: OpenAI/Azure: 350 gCO_2e/kWh; Google/GCP: 370 gCO_2e/kWh; Amazon (estimate across Oregon and Virginia grids): 287 gCO_2e/kWh; US-weighted default: 370 gCO_2e/kWh; Chinese grid average: 560 gCO_2e/kWh.

5.1 CogVideoX (Measured Anchor)

CogVideoX Low (2B model, 720×480, 8fps, 5s): Measured by Luccioni via CodeCarbon on H100. MIT applied 2× overhead. Total energy: 109,000 J for 5s = 21,800 J/s = 6.06 Wh/s. Location-based: 6.06 × 0.370 (US-weighted default grid intensity) = 2.24 gCO_2e/s. CEAF 0% (open-source, user-deployed). CEAF-adjusted: 2.24 gCO_2e/s. Range: 1.5–3.5 gCO_2e/s. Confidence: Medium.

CogVideoX High (5B model, 720×480, 16fps, 5s): Same measurement methodology. Total energy: 3,400,000 J for 5s = 680,000 J/s = 188.9 Wh/s. Location-based: 188.9 × 0.370 (US-weighted default grid intensity) = 69.9 gCO_2e/s. CEAF 0%. CEAF-adjusted: 69.9 gCO_2e/s. Range: 45–110 gCO_2e/s. Confidence: Medium.

The 31× difference between Low and High reflects both frame rate doubling (2×) and the substantially larger 5B model (≈15× additional compute). This demonstrates that model architecture dominates quality settings in determining video energy consumption.

5.2 Sora 2 Draft

OpenAI's draft-quality video mode. Generates at reduced resolution and diffusion steps for faster iteration. API generation time: 60–120s for a 5–10s clip. Runs on Azure + Oracle Stargate (H100/H200 class).

Step 1: Reference anchor. CogVideoX High = 188.9 Wh/s output; implied server power ~950W on a single H100 at ~60 min rendering time (Section 4.3).

Step 2: Rendering time. API generation time: 60–120s for a 5–10s clip. Rendering ratio: central ~12 rendering-seconds per output-second (range: 8–16 s/s).

Step 3: GPU cluster size assumption. Draft mode is designed for rapid iteration; OpenAI's Sora infrastructure at launch operated on Azure H100/H200 clusters. Central assumption: 16×H100 per inference request. Range: 8–32×. Basis: draft mode likely uses a smaller parallel cluster than full-quality Sora 2; industry precedent for mid-scale diffusion inference. This is the primary uncertainty.

Step 4: Energy from rendering time — Approach A. Central: 16 × 950W × 1.0 (hardware) × 12 ÷ 3,600 = 50.7 Wh/s. Range (8–32 GPUs, 8–16 s/s): 16.9–135 Wh/s. Note: Sora 2 Draft's output resolution is undisclosed; if it runs at reduced resolution, energy may be lower. If at full 1080p with fewer diffusion steps, the GPU count assumption dominates.

Step 5: Frame rate and resolution cross-check — Approach B. fps_ratio = 24/16 = 1.50. Resolution: Draft mode undisclosed; treating as approximately 720×480 (same as anchor), res_ratio ≈ 1.00. Quality factor = sqrt(1.50 × 1.00) = 1.22. Approach B: 188.9 × 1.22 × 1.0 = 230 Wh/s.

Step 6: Hardware correction = ×1.0 (H100/H200, same class as anchor).

Step 7: Central Wh/s estimate. Approach A: 50.7 Wh/s. Approach B: 230 Wh/s. These diverge by a factor of 4.5, driven entirely by the GPU cluster size assumption. A cluster of ~70×H100 would reconcile the two approaches; this is implausible for a draft mode. The divergence reflects a genuine methodological limitation. We retain Approach B (230 Wh/s) as the central figure and flag that Approach A suggests the estimate may be significantly overstated if the cluster is smaller than ~50 GPUs. Range: 17–230 Wh/s (incorporating Approach A lower bound). Central: 226.7 Wh/s..

Step 8: Location-based gCO_2e/s = 226.7 × 0.350 (OpenAI/Azure)= 79.3 gCO_2e/s.

Step 9: CEAF-adjusted (Azure+Stargate blended, CEAF 25%; see Section 7) = 79.3 × (1 − 0.25) = 59.5 gCO_2e/s. The 25% blended rate reflects the growing Oracle Stargate fraction with no clean energy disclosure.

Range: 5–130 gCO_2e/s location-based. CEAF-adjusted central: 62.9 gCO_2e/s (blended 25% CEAF). Confidence: Low.

Multi-generation workflow: Sora 2 Draft is intended for iterative refinement before committing to a full Sora 2 render, though it is not confirmed that most users follow this workflow in practice. See Section 4.3 for session-level accounting. A hypothetical workflow of 3 draft clips (5s each) before one 10s final render would add approximately 944 gCO_2e CEAF-adjusted (at 25% blended CEAF) on top of the final render cost.

5.3 Gemini Video (Veo)

Google's video generation model. Generates at up to 1080p (~24fps) on Google Cloud TPUs. No published energy data for Veo. API generation: 60–90s for an 8s clip.

Step 1: Reference anchor. CogVideoX High = 188.9 Wh/s output; implied server power ~950W on single H100 (Section 4.3).

Step 2: Rendering time. API generation time: 60–90s for an 8s clip. Rendering ratio: central ~9.4 rendering-seconds per output-second (range: 7.5–11.3 s/s).

Step 3: GPU cluster size assumption. Google runs Veo on TPU v5p/v6 pods. TPU pods typically comprise 256 chips in a full configuration; inference slices are smaller. Central assumption: 32 TPU-chip-equivalents per inference (in H100 efficiency-units; see Step 6). Range: 16–64. Basis: Google's published inference throughput and pod architecture. Highly uncertain; GPU/TPU cluster sizing for Veo is not disclosed.

Step 4: Energy from rendering time — Approach A. Before hardware correction: 32 × 950W × 9.4 ÷ 3,600 = 79.4 Wh/s. After TPU hardware correction ×0.5: 39.7 Wh/s. Range (16–64 TPU-equiv, 7.5–11.3 s/s, TPU factor 0.25–0.6): 8–81 Wh/s.

Step 5: Frame rate and resolution cross-check — Approach B. fps_ratio = 24/16 = 1.50. res_ratio = (1920×1080)/(720×480) = 6.00. Quality factor = sqrt(1.50 × 6.00) = sqrt(9.0) = 3.0. Approach B (before hardware): 188.9 × 3.0 = 566.7 Wh/s. After TPU ×0.5: 283.4 Wh/s.

Step 6: Hardware correction = ×0.5 (central). Google's TPU v5p/v6 are 2–3× more energy-efficient per FLOP than H100 (Patterson et al. 2022, Google Environmental Report 2025). Liquid-cooled custom silicon may achieve 3–4× (correction 0.25–0.33). Additionally, Veo's proprietary latent-space compression means pixel-count scaling (Approach B) likely overstates compute. Central: 0.5; range 0.25–0.6. Our Veo estimate likely overstates Google's actual energy.[8]

Step 7: Central Wh/s estimate. Approach A: 39.7 Wh/s. Approach B: 283.4 Wh/s. These diverge by a factor of 7. The divergence is larger than for Sora because Google's TPU efficiency cuts both estimates, but Approach B is further amplified by the high resolution quality factor (3.0×). Given that Veo's latent-space compression means Approach B likely over-estimates significantly, Approach A is arguably more credible here, though the TPU cluster size remains uncertain. We retain Approach B central (283.4 Wh/s) and flag that the true figure is likely substantially lower. Revised range: 8–284 Wh/s (Approach A lower bound to Approach B central).

Step 8: Location-based gCO_2e/s = 283.4 × 0.370 (Google) = 104.9 gCO_2e/s.

Step 9: CEAF-adjusted (Google, Tier 1, CEAF 66%) = 104.8 × (1 − 0.66) = 35.7 gCO_2e/s.

Range: 3–180 gCO_2e/s location-based (wider range incorporating Approach A lower bound). Confidence: Very Low.

5.4 Sora 2 (Full Quality)

OpenAI's flagship video generation model. Generates at 1080p, ~24fps with synchronised audio. Diffusion transformer architecture. API generation: 60–300s for a 10–25s clip. Runs on Azure + Oracle Stargate.

[8]Patterson, D., Gonzalez, J., Le, Q. et al. (2022) "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink." IEEE Computer 55(7). Google TPU efficiency also documented in Google 2025 Environmental Report: TPU v5p/Ironwood ~30x more power-efficient than Cloud TPU v2 (2018). https://sustainability.google/reports/google-2025-environmental-report/

Step 1: Reference anchor. CogVideoX High = 188.9 Wh/s output; implied server power ~950W on single H100 (Section 4.3).

Step 2: Rendering time. API generation time: 60–300s for a 10–25s clip. Rendering ratio: central ~18 rendering-seconds per output-second (range: 6–30 s/s). The wide range reflects variable clip length and server load.

Step 3: GPU cluster size assumption. Sora 2 full-quality is OpenAI's flagship video model, producing 1080p at 24fps with audio. This requires substantially more parallel compute than draft mode. Central assumption: 32×H100 per inference request. Range: 16–64×. Basis: scale consistent with OpenAI's Stargate infrastructure; full-quality generation requires more parallelism than draft. Primary uncertainty.

Step 4: Energy from rendering time — Approach A. Central: 32 × 950W × 1.0 × 18 ÷ 3,600 = 152.0 Wh/s. Range (16–64 GPUs, 6–30 s/s): 25.3–507 Wh/s.

Step 5: Frame rate and resolution cross-check — Approach B. fps_ratio = 24/16 = 1.50. res_ratio = (1920×1080)/(720×480) = 6.00. Quality factor = sqrt(9.0) = 3.0. Approach B: 188.9 × 3.0 × 1.0 = 566.7 Wh/s.

Step 6: Hardware correction = ×1.0 (H100/H200 class).

Step 7: Central Wh/s estimate. Approach A: 152.0 Wh/s. Approach B: 566.7 Wh/s. These diverge by a factor of 3.7. At 64×H100 (upper end of range), Approach A gives 304 Wh/s, approaching but not reaching Approach B. Geometric mean: sqrt(152 × 567) = 293.4 Wh/s. We update the central figure to the geometric mean: 293 Wh/s, which is better supported by the two independent approaches than the previous Approach-B-only figure.

Step 8: Location-based gCO_2e/s = 293 × 0.350 (OpenAI/Azure) = 102.6 gCO_2e/s.

Step 9: CEAF-adjusted (Azure+Stargate blended, CEAF 25%; see Section 7) = 102.6 × (1 − 0.25) = 76.9 gCO_2e/s. The 25% blended rate reflects the growing Oracle Stargate fraction with no clean energy disclosure.

Range: 7–315 gCO_2e/s CEAF-adjusted (25% blended). Location-based range: 9–420 gCO_2e/s. Confidence: Low.

Cross-check: SINTEF (2024) estimated the original Sora at ~19 Wh/s. Sora 2 full-quality would plausibly use 5–30× more compute, giving 95–570 Wh/s. Our revised central (293 Wh/s) falls within this range, providing additional support for the geometric mean approach.

5.5 Kling

Kuaishou's video model. Generates at 1080p, 30–48fps (Kling 2.5 Turbo). Diffusion-based transformer with 3D variational autoencoder. Runs on Chinese data centres (no clean energy disclosure). API generation: 2–3 minutes for a 5–10s clip on paid plans.

Step 1: Reference anchor. CogVideoX High = 188.9 Wh/s output; implied server power ~950W on single H100 (Section 4.3).

Step 2: Rendering time. API generation time: 2–3 minutes for a 5–10s clip. Rendering ratio: central ~24 rendering-seconds per output-second (range: 12–36 s/s). Kling's relatively long generation time is consistent with its higher frame rate and complex 3D variational autoencoder architecture.

Step 3: GPU cluster size assumption. Kling runs on Kuaishou's Chinese data centres; hardware specifications are unpublished. Central assumption: 16×H100-equivalent per inference. Range: 8–32×. Basis: comparable scale to Runway and mid-tier commercial providers. This is highly uncertain; Kuaishou has claimed 62% lower compute costs for Kling 2.5 Turbo, suggesting significant inference optimisation that may imply a smaller or more efficient cluster than our central assumption.[9]

[9]Kuaishou Technology, "Kling AI Launches 2.5 Turbo Video Model," 26 September 2025. Reports 62% lower credit cost (105 vs 280 credits for 10s video) and 60% faster generation vs Kling 2.1 Master.
https://ir.kuaishou.com/news-releases/news-release-details/kling-ai-launches-25-turbo-video-model-industry-leading

Step 4: Energy from rendering time — Approach A. Central: 16 × 950W × 1.0 × 24 ÷ 3,600 = 101.3 Wh/s. Range (8–32 GPUs, 12–36 s/s): 25.3–304 Wh/s.

Step 5: Frame rate and resolution cross-check — Approach B. fps_ratio = 30/16 = 1.875. res_ratio = (1920×1080)/(720×480) = 6.00. Quality factor = sqrt(1.875 × 6.00) = sqrt(11.25) = 3.35. Approach B: 188.9 × 3.35 × 1.0 = 632.8 Wh/s.

Step 6: Hardware correction = ×1.0 (infrastructure unknown; Chinese data centres, no published hardware specifications; unknown direction of error).

Step 7: Central Wh/s estimate. Approach A: 101.3 Wh/s. Approach B: 632.8 Wh/s. Divergence factor: 6.2×. Geometric mean: sqrt(101 × 633) = 252.9 Wh/s. We update the central to the geometric mean: 253 Wh/s. Location-based: 253 × 0.560 (Chinese grid) = 141.7 gCO_2e/s. Note: Kuaishou's claimed 62% compute reduction for Kling 2.5 Turbo and its 3D VAE architecture both suggest Approach B may overstate the true energy; the geometric mean is therefore likely still conservative. The estimate is flagged as an upper-bound proxy.

Step 8: Location-based gCO_2e/s = 141.7 (geometric mean of Approach A and B).

Step 9: CEAF-adjusted (Tier 3, CEAF 0%) = 141.7 × (1 − 0.00) = 141.7 gCO_2e/s.

Range: 9–470 gCO_2e/s location-based. Confidence: Very Low.

Note: Kling's higher frame rate (30fps vs 24fps for Sora/Veo) contributes to the higher Approach B quality factor. Kuaishou's claim of 62% lower compute costs and the 3D VAE architecture with proprietary spatiotemporal compression both suggest the true figure may be substantially below our central; the geometric mean is intended to partially address Approach B's known tendency to overstate for architecturally distinct models.

5.6 Runway Gen-4 / Gen-4.5

Runway's current production video models. Gen-4 (March 2025) generates at native 1080p, 24fps, with 4K upscaling available. Gen-4.5 (December 2025) is the latest variant with improved motion fidelity and prompt adherence, ranked #1 on the Artificial Analysis text-to-video leaderboard (1,247 Elo). Both output at higher resolution than Gen-3 Alpha (which was 1280×768). Infrastructure undisclosed. Gen-4 Turbo generates a 10s clip in ~30 seconds; standard Gen-4 takes 45–90s for a 5s clip.[10]

Step 1: Reference anchor. CogVideoX High = 188.9 Wh/s output; implied server power ~950W on single H100 (Section 4.3).

Step 2: Rendering time. Standard Gen-4: 45–90s for a 5s clip. Gen-4 Turbo: ~30s for a 10s clip. Rendering ratio (standard): central ~13.5 rendering-seconds per output-second (range: 9–18 s/s). Gen-4 Turbo rendering ratio: ~3 s/s — a 4.5× reduction versus standard, providing strong evidence of inference optimisation.

Step 3: GPU cluster size assumption. Runway's infrastructure is undisclosed. Central assumption: 16×H100-equivalent per inference (standard Gen-4). Range: 8–32×. Gen-4 Turbo may use a similar cluster at higher utilisation or a smaller cluster, explaining its speed. The ×1.0 hardware correction does not assume Runway is inefficient; the unknown direction of error is acknowledged.

Step 4: Energy from rendering time — Approach A. Central (standard Gen-4): 16 × 950W × 1.0 × 13.5 ÷ 3,600 = 57.0 Wh/s. Gen-4 Turbo (3 s/s rendering ratio, same GPU count assumption): 16 × 950W × 3 ÷ 3,600 = 12.7 Wh/s. Range (standard, 8–32 GPUs, 9–18 s/s): 19–152 Wh/s.

Step 5: Frame rate and resolution cross-check — Approach B. fps_ratio = 24/16 = 1.50. res_ratio = (1920×1080)/(720×480) = 6.00. Quality factor = sqrt(9.0) = 3.0. Approach B: 188.9 × 3.0 × 1.0 = 566.7 Wh/s. Note: Gen-4's credit cost increased from 10 to 12 credits/sec vs Gen-3 Alpha, consistent with higher compute per second of output.

Step 6: Hardware correction = ×1.0 (infrastructure unknown; ×1.0 represents unknown direction of error, not assumed inefficiency).

[10]Artificial Analysis Text-to-Video Arena leaderboard (March 2026). Runway Gen-4.5 ranked #1 at 1,247 Elo. https://artificialanalysis.ai/text-to-video/arena

Step 7: Central Wh/s estimate. Approach A (standard Gen-4): 57.0 Wh/s. Approach B: 566.7 Wh/s. Divergence factor: 9.9×. Geometric mean: sqrt(57 × 567) = 179.7 Wh/s. We update the central to the geometric mean: 180 Wh/s. Location-based: 180 × 0.370 (US-weighted default) = 66.6 gCO_2e/s. Gen-4 Turbo's Approach A figure (12.7 Wh/s) demonstrates that for the Turbo variant the true energy may be far lower; this is consistent with Turbo's 5× speed improvement and half the credit cost.

Step 8: Location-based gCO_2e/s = 66.6 (geometric mean of Approach A and B).

Step 9: CEAF-adjusted (Tier 3, CEAF 0%) = 66.6 × (1 − 0.00) = 66.6 gCO_2e/s.

Range: 7–420 gCO_2e/s location-based. Confidence: Very Low.

The Runway estimate (66.6 gCO_2e/s) reflects the geometric mean, which partially corrects Approach B's tendency to overstate for models whose inference architecture differs significantly from CogVideoX. The credit-cost signal (12 credits/sec for Gen-4 vs 10 for Gen-3 Alpha) remains a useful corroborating data point for the relative ordering of Runway vs older models.

6. Video Quality Multipliers

This section provides video quality multipliers as reference material. The executive summary figures are derived from the energy-per-second scaling methodology (Section 4.3), not from these multipliers.

Quality multipliers scale estimates relative to a 720p 16fps baseline. Video energy scales super-linearly with resolution and frame rate.

Quality	Multiplier	Pixel×Frame Ratio	Assessment
480p 8fps	0.30×	0.22×	Slightly generous vs raw ratio. Directionally correct.
720p 16fps	1.00×	1.00×	Baseline. Matches CogVideoX High measurement conditions.
1080p 24fps	2.80×	3.38×	Below raw ratio; reflects hardware optimisations at common resolutions.
4K 30fps	8.00×	16.88×	Extrapolated. Raw ratio 16.9×; heavily discounted. Low confidence.

Table 3: Video quality multipliers. The measured CogVideoX low-to-high ratio (31×) is far larger than these multipliers predict, because that "quality" change involved a completely different model, not just resolution/framerate scaling. These multipliers should only be applied within a single model.

7. Location-Based vs Clean Energy Adjustment Factor (CEAF)-Adjusted Accounting

This document reports estimates in two layers, following the GHG Protocol Scope 2 framework:

Location-based (gross): Energy consumed × grid carbon intensity at the data centre location. This reflects the physical carbon impact of the electricity drawn from the grid. It is the primary metric for Scope 2 transparency reporting and is required by IFRS S2.

CEAF-adjusted: Location-based × (1 − CEAF). The Clean Energy Adjustment Factor (CEAF) is a discount applied to reflect the provider's verified clean energy investment. It acknowledges that providers purchasing or generating renewable electricity reduce the marginal grid emissions attributable to their operations. CEAF-adjusted figures are the primary decision-support metric in the InferenceCarbon calculator, as they reward genuine clean energy procurement and better reflect the net carbon impact for comparative purposes.

CEAF = Location-based × (1 − CEAF)

CEAF tiers: Tier 1 — Verified hourly Carbon-Free Energy (CFE): Full percentage applied. Tier 1.5 — Verified annual renewable match, no hourly data: 50% of 100%. Tier 2 — Annual Renewable Energy Certificates (RECs), weaker verification: 50% of claimed percentage. Tier 3 — No disclosure: CEAF = 0%.

OpenAI / Oracle Stargate CEAF: A worked example of the blended CEAF problem.

OpenAI historically carried a 50% CEAF derived entirely from Microsoft Azure's 100% annual renewable match (Tier 1.5). However, OpenAI's Stargate partnership with Oracle (announced January 2025, Abilene TX campus operational from mid-2025) changes this materially.[11]

Oracle has not published any clean energy commitments for the Stargate campus that we have been able to identify. The Abilene, Texas facility sits on the ERCOT grid, which as of 2024 runs at approximately 380 gCO_2e/kWh — close to the US average but with no renewable match applied. The fraction of Sora inference running on Stargate vs Azure is not disclosed by OpenAI.

Blended CEAF calculation: If we assume 50% of Sora inference is on Azure (CEAF 50%) and 50% on Stargate (CEAF 0%), the blended CEAF = (0.50 × 50%) + (0.50 × 0%) = 25%. If the Stargate fraction is higher — which is likely given it is the primary expansion infrastructure — the true blended CEAF may be below 25%. We adopt 25% as the central estimate.

Tripwire conditions for CEAF revision:

Upgrade to 50%: Oracle publishes a clean energy commitment or Power Purchase Agreement for the Stargate Abilene campus covering ≥50% of consumption; OR OpenAI publishes its own infrastructure carbon disclosure showing ≥50% clean energy across the blended fleet.

Upgrade to 66%+: Oracle achieves hourly CFE matching for Stargate; OR Google-equivalent disclosure from OpenAI of hourly matched clean energy across all inference infrastructure.

Downgrade to 0%: Evidence emerges that >90% of Sora inference runs on Stargate with no clean energy commitment; OR OpenAI's infrastructure carbon disclosure shows <10% clean energy fleet-wide.

Maintain at 25%: No new disclosure from Oracle or OpenAI; Stargate fraction unknown but material.

The CEAF-adjusted figure is presented as the primary estimate in this document. The location-based gross figure is retained as the physical baseline and is available for organisations requiring unadjusted Scope 2 reporting.

8. Confidence Assessment

Element	Confidence	Rationale
CogVideoX energy (Luccioni/MIT)	Medium	GPU directly measured via CodeCarbon. 2× overhead standard.
CogVideoX grid conversion	High	370 gCO_2e/kWh (US-weighted) is well-established.
Sora 2 estimates	Low	Scaled from CogVideoX High via sqrt(quality ratio). No direct measurement.
Kling / Runway Gen-4	Very Low	No published data of any kind. Engineering guesses.
Gemini Video (Veo)	Very Low	Google has not published Veo energy data.
Quality multipliers	Low	Directionally correct but 4K value is extrapolated.
CEAF values	Medium-Low	Provider-level; actual video inference routing unknown.

Table 4: Confidence assessment.

9. Limitations

1. **Only one model family measured.** CogVideoX is the only video model with published energy data.

2. **31× quality sensitivity.** The CogVideoX low-to-high ratio shows that model architecture dominates quality settings. Multipliers based on resolution/framerate alone are inadequate.

[11]OpenAI, Oracle, and SoftBank announced the Stargate AI infrastructure project at the White House on 21 January 2025. Flagship campus in Abilene, Texas operational from mid-2025 on Oracle Cloud Infrastructure. https://openai.com/index/five-new-stargate-sites/

3. **Sora 2 estimates are speculative.** Estimated via quality-ratio scaling from CogVideoX High (Section 4.3). SINTEF's "1,800 image generations" is used as a cross-check only (Section 5.4), not as the primary derivation.

4. **Kling, Runway, and Gemini Video have zero data.** All estimates are engineering guesses.

5. **Video generation is evolving rapidly.** Models released months apart may differ by 10× in efficiency. These estimates will age quickly.

6. **The sqrt() quality-ratio moderation is a heuristic. Commercial models may be significantly more or less energy-efficient than predicted.** Excludes training (Sora training estimated at 100,000 GPU-hours), embodied emissions, and supply chain.

7. **CEAF uncertainty for OpenAI.** Sora 2 runs on Azure + Oracle Stargate. The Stargate fraction is unknown but material; Stargate has no clean energy disclosure. CEAF set at 25% blended. See Section 7 for tripwire conditions.

10. Response to Adversarial Review

This document was reviewed by Gemini 2.5 Pro (thinking) from dual perspectives at each provider: a PhD data centre engineer and a senior sustainability executive.

10.1 Technical critiques accepted

Google TPU factor too conservative: The reviewer argued that liquid-cooled TPU v5p/v6 may achieve 3–4× efficiency vs H100. Accepted. We have widened the TPU range to 0.25–0.6 and noted that our Veo estimate likely overstates Google's actual energy. Latent-space compression caveat also added.

Latent space optimisation: The reviewer correctly noted that modern video models operate in compressed latent space, meaning pixel-count scaling overstates the actual compute ratio. Accepted. This is what our sqrt() moderation attempts to capture, but we have added an explicit caveat in Section 4.3 acknowledging that sqrt() may still be insufficient for highly optimised commercial models.

Kling Turbo efficiency: Kuaishou claims 62% lower compute costs with Kling 2.5 Turbo. Our central estimate does not capture this optimisation. Accepted — caveat strengthened; estimate flagged as an upper-bound proxy. Kling's 3D variational autoencoder architecture also noted as a cross-architecture limitation.

Commercial PUE < 2×: Purpose-built facilities (Oracle Stargate, Google's TPU pods) likely achieve PUE below 1.2, meaning the 2× overhead baked into the CogVideoX anchor may overstate commercial infrastructure by 20–30%. Partially accepted — caveat added to Section 2, noting all estimates inherit this potentially conservative assumption.

Runway model update: Following this critique, we updated Section 5.6 from Gen-3 Alpha to Gen-4/Gen-4.5 (see Section 10.4 for detailed analysis). Counterintuitively, the estimate increased because Gen-4/4.5 outputs at higher resolution (1080p vs 1280×768). Runway's ×1.0 hardware factor clarified as unknown direction, not assumed inefficiency.

10.2 Sustainability critiques

Microsoft/OpenAI CEAF "penalty" (now 25% blended): Respectfully rejected. Consistent with the full InferenceCarbon series. Annual renewable matching is an accounting methodology, not a guarantee of carbon-free electricity every hour. The blended 25% rate reflects the growing Stargate fraction with no clean energy disclosure. We commit to upgrading when fleet-wide hourly CFE data is published.

Google grid siting: The reviewer argued that Google routes video generation to high-CFE regions. Noted but rejected for consistency. We cannot know which data centre handles any given API call. The US-weighted average is applied consistently across all providers in the series.

Google "Very Low" confidence + 66% CEAF: The reviewer argued this undermines the transparency incentive. Clarification: "Very Low" refers to the energy estimate confidence, not the CEAF confidence.

Google's CEAF is rated Medium confidence (Table 4) based on auditable hourly data. The energy estimate for Veo is Very Low because Google has not published any Veo-specific energy data.

Kling 0% CEAF for Chinese infrastructure: The reviewer argued that Chinese DC providers are investing in solar and wind. Noted but maintained. The CEAF rewards disclosure, not investment intent. We acknowledge that major Chinese providers are procuring renewable energy; however, without published hourly or annual matching data, we cannot verify clean energy claims. If Kuaishou discloses its energy sourcing, we will assign an appropriate CEAF.

Scope 3 omission: Valid observation. The embodied carbon of H100/H200 clusters is significant. However, Scope 3 is explicitly out of scope for this document, as stated in the scope notice. A future InferenceCarbon paper on lifecycle assessment may address this.

10.3 Net impact

Central estimates for Sora 2, Kling, and Runway have been revised using the geometric mean of Approach A and B (see Section 4.3). Veo's range widens downward (TPU factor + latent optimisation). Kling's central estimate flagged as an upper-bound proxy (Turbo gains + architecture differences). All commercial estimates are flagged as likely conservative due to PUE < 2× in purpose-built facilities. The reviewer's core critique — that sqrt(quality_ratio) applied cross-architecture is speculative — is fair. We accept this as a fundamental limitation of the only available approach when no provider publishes video generation energy data.

10.4 Adversarial review: Runway Gen-4/4.5 update

The initial Gemini review argued that Gen-3 Alpha was obsolete and that newer models (Gen-4, Gen-4.5) would likely be "more efficient." We updated Section 5.6 to assess Gen-4/4.5 instead. The initial Approach-B-only estimate was 209.7 gCO_2e/s; applying the hybrid methodology (geometric mean of Approach A and B) yields a revised central of 66.6 gCO_2e/s.

Anticipated Runway engineer critique: "Gen-4 Turbo is 5× faster at half the credit cost. Your estimate penalises us for delivering higher quality, while ignoring the efficiency gains Turbo demonstrates. The true energy-per-second for Gen-4 Turbo is likely far below your central figure of 66.6 gCO_2e/s."

Response: Fair. Gen-4 Turbo generates a 10s clip in ~30 seconds at 5 credits/sec (vs standard Gen-4 at 12 credits/sec), representing significant inference optimisation. However, credit pricing reflects business positioning, not just compute — the same caveat we apply to OpenAI TTS HD in the audio paper. We note the Turbo variant as evidence that our central estimate is likely conservative for optimised deployments, but retain the standard Gen-4 derivation as the baseline because Turbo may trade quality for speed.

Anticipated Runway sustainability critique: "With 0% CEAF, Runway Gen-4/4.5's CEAF-adjusted figure (66.6 gCO_2e/s) receives no clean energy discount, while Sora 2 (81.3 gCO_2e/s at 25% CEAF) does. The difference in treatment reflects disclosure, not a physical measurement."

Response: This observation illustrates the CEAF framework's central incentive. Runway's CEAF-adjusted estimate (66.6 gCO_2e/s) and Sora 2's (81.3 gCO_2e/s) now differ less on a location-based basis (66.6 vs 108.4) due to the geometric mean revision, but the CEAF gap remains: Runway carries 0% CEAF (undisclosed infrastructure) while OpenAI carries 25% (blended Azure + Stargate). Disclosure of clean energy investment is rewarded. If Runway publishes its infrastructure and energy sourcing, we will assign an appropriate CEAF.

Net impact: Runway estimate revised to 66.6 gCO_2e/s (geometric mean methodology). Gen-4 Turbo noted as evidence that optimised inference may be substantially lower. The Runway/Sora 2 CEAF divergence is documented as a feature of the framework, not a flaw.

11. Key Sources

Primary empirical anchor: Luccioni/MIT Technology Review (May 2025) CogVideoX measurements. SINTEF (2024) provides directional context for Sora but is not a measurement. Chung & Chowdhury

(2026) provides framework for understanding video energy scaling. All sources cited in footnotes with URLs. Additional sources referenced in text: MIT Technology Review methodology companion (technologyreview.com/2025/05/20/1116331/); Google 2025 Environmental Report (sustainability.google); Microsoft carbon negative milestone (blogs.microsoft.com/blog/2026/02/18/); OpenAI Sora 2 announcement (openai.com/index/sora-2/); Runway Gen-4.5 (Artificial Analysis text-to-video leaderboard, artificialanalysis.ai); Kuaishou Kling 2.5 Turbo release notes (klingai.com).