Green Quality Distance: Characterizing Energy and Quality Trade-Off in Open Text-to-Image and Text-to-Video Models

Anonymous authors
Paper under double-blind review

Abstract

Advances in generative technologies have led to the development of models capable of creating high-quality images and videos from text. In turn, the environmental impact of deployment and usage is a growing concern in research and industry. These models are commonly evaluated using qualitative metrics related to perceptual quality and prompt adherence. Recent research focused on measuring the consumed energy or the emitted CO_2 during training and inference. The notion of *Intelligence per Watt* (Saad-Falcon et al., 2026) enables a rating of Large Language Models based on their quality (e.g. intelligence) relative to the required power (e.g. watts). This metric cannot, however, be readily applied to image and video models. We propose a novel formulation allowing us to assess both the visual quality and environmental impact of generative image and video models for a given inference configuration: the **Green Quality Distance**. Our metric emphasizes the environmental impact of models by gathering data on energy consumption, abiotic resource usage, global warming potential, and water use. By benchmarking configurations (e.g. diffusion steps, GPU power cap) of open-source state-of-the-art models, we show that the right configuration reaches a given quality bar at a fraction of the energy cost, exploiting the full potential of existing models while reducing resource waste.

1 Introduction

Over the past few years, generative models for images and videos have evolved from research demonstrations to widely deployed creative tools. State of the art diffusion and transformer models are now capable of synthesizing highly realistic images, editing existing photographs, generating images and videos from text, and assisting professionals in creative workflows. Their adoption has been incredibly fast: hundreds of millions of users now interact with generative AI systems each week (Liu & Wang, 2026), while Adobe (2025) indicates that more than 85% of professional creators already incorporate generative AI into their workflows.

This unprecedented democratization of visual content generation also implies an equally unprecedented demand for computation. Unlike traditional editing pipelines, modern diffusion and transformer-based generators require substantial GPU resources during training and inference. As the frequency of image and video generation continues to grow, quantifying, communicating and reducing the associated environmental footprint becomes increasingly important for the sustainable development of generative AI.

Platforms and software featuring proprietary opaque models generally expose few parameters, without sharing knowledge of their associated environmental impact. The number of diffusion steps is often used as a proxy for final quality. For open-weight models, users may stay on the model's default settings without understanding their impact. In this paper, we aim at solving this problem by providing a novel formulation, the **Green Quality Distance**, to evaluate a model's quality against its environmental impact, as well as a computational framework that can be used by model providers for both improving their transparency and providing user levers to reduce the environmental footprint of their creative workflows.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We evaluate 18 open-weight image and video generative models using our **Green Quality Distance**, to assess their performance (both qualitatively and resource-wise) against changes in denoising steps, resolution, and GPU power capping, totaling 140 configurations. We demonstrate that for the most efficient models (for instance *FLUX.2 Klein 9b*), changing the inference settings can save up to 81% of energy while maintaining a similar quality. More specifically, our contributions are the following:

1. **Green Quality Distance.** We introduce a novel metric that jointly ranks image and video generative models by quality **and** environmental cost, expressed as a single distance to an ideal (maximum quality and zero impact) operating point.
2. **Benchmarks.** Using this metric, we benchmark a range of open-weight image and video models, quantifying how quality and energy cost shift with three inference settings: number of denoising steps, output resolution, and GPU power capping.
3. **Configuration optimization.** As a direct application of our work, we propose a model-agnostic methodology for identifying per-model inference configurations that preserve most of a model's quality at a fraction of its energy cost.

2 Related Work

Our work lies at the intersection of two research topics: the evaluation and benchmarking of generative models, and the environmental accounting of generative AI technologies.

GenAI image and video models evaluation. Quantitative evaluation of **image models** was introduced with likelihood and inception based metrics (Salimans et al., 2016). The *Fréchet Inception Distance* (Heusel et al., 2017) and *Kernel Inception Distance* (Bińkowski et al., 2018) compare the statistical distribution of features between real and generated samples. With the rise of text-to-image models based on latent diffusion (Rombach et al., 2022), diffusion transformers (Peebles & Xie, 2023) became the standard for image generation, and evaluation shifted towards conditional metrics. CLIPScore (Hessel et al., 2021) uses CLIP embeddings (Radford et al., 2021) to measure image alignment to its prompt. *GenEval* (Ghosh et al., 2023) adds object-centric and composition criteria by checking explicit attributes such as object counting and color attribution. We rely on these metrics for the quality axis of our formulation. Other metrics were developed for evaluating the quality of **videos**, such as the Fréchet Video Distance (FVD) (Unterthiner et al., 2019). As noted by Ge et al. (2024), FVD shows a bias towards the quality of individual frames, focusing less on temporal coherency. VBench (Huang et al., 2024; 2025; Zheng et al., 2025) provides a more complete evaluation benchmark, adding criteria such as motion smoothness and temporal style. VideoPhy (Bansal et al., 2025) and Physics-IQ (Motamed et al., 2026) propose two evaluation frameworks to measure the compliance of generative videos with physical laws. We use VBench in our video benchmark but other metrics could easily be integrated.

Despite being instrumental for tracking GenAI progress, these metrics focus solely on quality and often present best performers on static leaderboards under fixed parameter sets. In contrast, we explore the effect of inference parameters in the quality-energy trade-off, identifying both non-necessary resource waste and truly efficient high-quality configurations.

AI models impact. Concerns were raised regarding the environmental impact of GenAI given its massive development, starting with Strubell et al. (2019) for LLMs. The GreenAI agenda (Schwartz et al., 2020) calls for efficiency to be treated as a major evaluation criterion alongside accuracy. Luccioni et al. (2022) extend this perspective to the full development cycle of a 176B-parameter LLM, showing the considerable carbon footprint of embodied emissions and idle infrastructure. For inference, Luccioni et al. (2024) compare a wide range of AI-based tasks and show that image generation is orders of magnitude more resource-intensive than other pipelines. These measures rely on tools such as CodeCarbon (Courty et al., 2024), for the carbon footprint estimation of code execution, and NVML (NVIDIA Corporation, 2026), for a mW telemetry of the per-device GPU power, enabling nearly exact energy consumption measurement. Recently, Delavande et al. (2025) raise concerns about video models, finding that inference of a few-seconds video can require more than 100Wh.

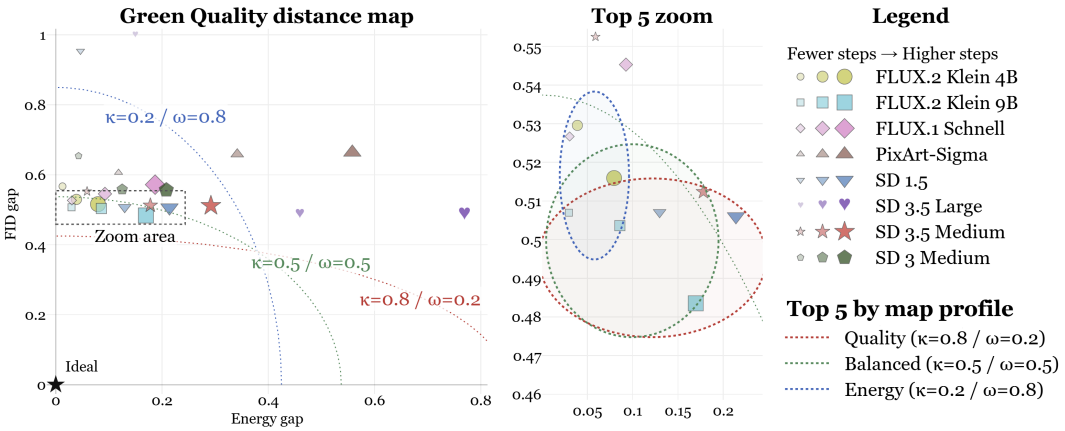


Figure 1: Green Quality Distance: Distance between benchmarked configurations and ideal configuration $\star$ for image generation at 512^2 resolution. Ellipses are iso-distances weighted by κ and ω, showing how the distance to $\star$ and the Top 5 evolve for different parameters.

Most of these works focus on benchmarking either quality or environmental cost, but Saad-Falcon et al. (2026) evaluates both simultaneously with *Intelligence per Watt*, capturing the useful capability an LLM delivers per unit of energy, enabling comparisons across quality-efficiency trade-offs. However, this formulation, designed for LLMs, leads to interpretation and scale limitations when applied to images, limitations that the Green Quality Distance addresses.

Infrastructure impact. Beyond direct GPU energy usage, several recent works enlarge the measurement boundary. BoaviztAPI (Simon et al., 2024) and EcoLogits (2026) provide models aggregating the impacts of hardware components and infrastructure overhead, reporting indicators such as Abiotic Depletion Potential (ADP) and Global Warming Potential (GWP) rather than energy or carbon alone. ADP tracks the depletion of non-renewable resources (mineral, metal, fossil). GWP quantifies the contribution of greenhouse-gas emissions to climate change. Falk et al. (2026) go further with a cradle-to-grave lifecycle assessment of the Nvidia A100 GPU, showing the impact of GPU manufacturing on mineral depletion and pollution. The Hubblo GPU-calculator (Lees-Perasso et al., 2026) extends this to other cards (H100, P100, RTX Titan) used for training and inference of GenAI models.

Building on these works, our formulation is flexible in terms of qualitative considerations and environmental costs. Depending on the desired focus, our metric allows considering energy alone or a broader scope including embodied impacts (ADP, GWP, Water Use).

3 Method

We aim at shifting the focus of generative model evaluation from a single-axis ranking to a multi-criteria comparison accounting for quality, energy usage, and environmental impact. The core concept involves evaluating models along with their inference configurations rather than by name alone, to identify good trade-offs. A model can occupy different positions on the quality-energy plane depending on resolution, denoising budget, or GPU power capping.

We define an inference configuration as $c = (m, r, s, g)$, with m the generative model, r the output resolution (and number of frames and frame rate for videos), s the number of denoising steps, g the guidance setting. For each configuration c, the framework records:

$$\mathcal{M}(c) = \big(Q(c), E(c), L(c), I(c)\big), \tag{1}$$

where $Q(c)$ denotes the quality metric(s), $E(c)$ the measured or estimated GPU energy, $L(c)$ the generation latency, and $I(c)$ the environmental impact estimate.

3.1 Environmental impact measurement protocol

Each evaluated configuration uses fixed prompts and seeds. FID, KID and CLIPScore benchmarks rely on COCO2017 captions (Lin et al., 2014), a dataset widely used for object detection, segmentation and recognition. GenEval and VBench use their respective official prompt suites. We start an NVIDIA Management Library (NVML) sampler, synchronize the GPU, and run the generation pipeline. Energy is computed by integrating the measured GPU power over the generation window $[t_0, t_1]$: $E(c) = \Omega \int_{t_0}^{t_1} P_{\mathrm{GPU}}(t)\, dt$, where $P_{\mathrm{GPU}}(t)$ is the instantaneous GPU power reported by NVML, and $\Omega = \frac{1}{3.6\times10^6}$ converts joules to kWh. For multi-GPU runs, the integral is summed across devices. We store the generated artifact, average power, peak power usage, and latency $L(c) = t_1 - t_0$. We report three environmental indicators, $j \in \{\mathrm{ADP}, \mathrm{GWP}, \mathrm{WU}\}$: abiotic depletion potential, global warming potential, and water use. Each has an *embodied* contribution I^{emb} driven by latency, and an *operational* contribution I^{op} driven by energy, summing to $I_j(c) = I_j^{\mathrm{emb}}(c) + I_j^{\mathrm{op}}(c)$.

Embodied impacts come from the hardware lifecycle and are amortized over the GPU's active lifespan T_{life}, reported as 3 to 6 years by Falk et al. (2026) and Luccioni et al. (2022). A single generation is allocated the fraction of this lifespan it occupies, $\frac{L(c)}{T_{\mathrm{life}}}$, applied to the GPU's total embodied impact $I^{\mathrm{emb}}_{\mathrm{life},j}$ for indicator j, (aggregated over manufacturing, transport, and end of life), with factors taken from Hubblo's `GPU-calculator` inventory (Lees-Perasso et al., 2026). For the WU indicator, we follow the same methodology as Falk et al. (2026) and exclude the end-of-life stage as the background database is subject to errors in this particular criterion. The per-generation embodied impact is thus:

$$I_j^{\mathrm{emb}}(c) = \frac{L(c)}{T_{\mathrm{life}}} I^{\mathrm{emb}}_{\mathrm{life},j} \tag{2}$$

Operational impacts arise from the electricity consumed during generation and are therefore proportional to the consumed energy $E(c)$ expressed in kWh. For GWP, energy is weighted by the carbon intensity $\varepsilon_{\mathrm{grid}}$ that accounts for the regional electricity mix, and for WU, by the on-site Water Usage Effectiveness (WUE). In our model, ADP does not include an operational term as we restrict our study to GPU materials, which are fully covered in the embodied impacts. The per-generation operational impact is thus:

$$I_j^{\mathrm{op}}(c) = E(c)\, \varepsilon_j^{\mathrm{op}}, \qquad \varepsilon^{\mathrm{op}}_{\mathrm{ADP}} = 0, \quad \varepsilon^{\mathrm{op}}_{\mathrm{GWP}} = \varepsilon_{\mathrm{grid}}, \quad \varepsilon^{\mathrm{op}}_{\mathrm{WU}} = \mathrm{WUE}. \tag{3}$$

3.2 Quality metrics

We consider quality as a multidimensional concept rather than a single property of a model. For image generation, this means accounting for visual realism and diversity of the generated distribution, as well as the prompt adherence. For videos, we add the temporal dimension such as motion and frames consistency. Importantly, our method is not tied to a particular set of metrics: $Q(c)$ may incorporate additional or not-yet-created evaluation criteria.

We propose a benchmark with FID to assess text-to-image models' quality, on a subset of 10k samples from COCO2017, complemented with CLIPScore (Hessel et al., 2021) to measure text-image alignment, as well as GenEval (Ghosh et al., 2023) for compositional coherence. For text-to-video, we use VBench (Huang et al., 2024) global and dimensional scores which separate image and motion quality, semantic alignment, and temporal consistency.

3.3 Green Quality Distance

Rather than ranking configurations by an unbounded efficiency ratio, we score each one by its *distance to an ideal*: a system with the best quality and lowest energy. Both quality $Q(c)$ and energy $E(c)$ are mapped onto a common, dimensionless *gap to the ideal* using fixed bounds. For minimized metrics (energy, FID) with a fixed reference cap (E_{lim}, Q_{lim}), and for a maximized quality metric (GenEval, VBench) with known maximum Q_{lim}, respectively:

$$d_E(c) = \frac{E(c)}{E_{\mathrm{lim}}}, \quad d_Q(c) = \frac{Q(c)}{Q_{\mathrm{lim}}} \qquad \text{or} \qquad d_Q(c) = 1 - \frac{Q(c)}{Q_{\mathrm{lim}}} \in [0, 1], \tag{4}$$

We introduce the **Green Quality Distance** (GQD), which scores configurations by their weighted Euclidean distance to the ideal $(d_Q, d_E) = (0, 0)$:

$$GQD_{\kappa,\omega}(c) = \sqrt{\kappa d_Q(c)^2 + \omega d_E(c)^2}, \tag{5}$$

with $\kappa, \omega \geq 0$ and $\kappa + \omega = 1$. Parameter κ controls the importance of quality and ω controls the penalty assigned to energy. A smaller distance implies a model closer to the ideal configuration for a given (κ, ω), with $\kappa = \omega = 0.5$ being the balanced setting. The energy term d_E could be replaced by latency d_L or impact d_I in Eq. 4 and 5. We provide theoretical, statistical and reproducibility elements in Appendix B.

4 Benchmarks: quality versus energy

4.1 Image Generation Benchmarks



Figure 2: Qualitative comparison of the denoising budget impact for *FLUX.2 klein 4B*, *9B*, and *Stable Diffusion 3.5 Large* and *3 Medium*, with aggregated FID scores. The first two are distilled models and are thus more robust to a low amount of steps. While *Flux.2 Klein 4B* may look good at 2 steps, artifacts can be seen around the trees (red circle) (More results in Fig. 11 and Fig. 12).

We benchmark 11 open-weight text-to-image (t2i) models on a subset of 10k samples from COCO2017. We measure energy, FID, CLIPScore, and Green Quality Distance at 512^2 and 1024^2 resolutions (when possible) with 3 diffusion steps count, leading to 48 configurations. Per image, the consumed energy using our protocol ranges from 0.013 to 4.55$\tilde{\text{W}}$h.

While the impact of resolution is predictable and monotonic, the denoising budget offers a lever for trading quality and impact. While energy consumption evolves linearly, the models are not equally robust at low denoising steps (Fig. 2, 11, 12 and Tab. 1). Distilled models are made to run at lower regimes, and may provide equivalent results at 12 steps instead of 50 for base models. Still, distilled or not, some models give similar results at a fraction of the recommended number of steps, while others strongly degrade the result.

Figure 1 illustrates our *distance to the ideal* formulation. The quality-energy trade-off is not linear: some configurations achieve marginal quality gains at large energetic costs, while others achieve similar quality at lower energy. This confirms the efficiency of our formulation for finding the most efficient configurations and using a model at its best potential.

Table 1: Image benchmark (extract of Tab. 4) with 3 inference step counts at 1024^2 pixels. Color bars indicate the per-column ranking (larger is better). We report CLIPScore, FID, GenEval, energy consumption, and our GQD for FID and GenEval, with 3 weightings: Balanced ($\kappa=\omega=0.5$), Quality-first ($\kappa=0.8$, $\omega=0.2$), Energy-first ($\kappa=0.2$, $\omega=0.8$)

Model	step #	energy ↓	CLIP score↑	FID ↓	GQD B↓	GQD Q↓	GQD E↓	Gen Eval↑	GQD B↓	GQD Q↓	GQD E↓
SDXL	10	0.13	31.44	28.60	0.54	0.68	0.34	0.48	0.37	0.46	0.23
SDXL	30	0.36	31.44	29.33	0.55	0.70	0.36	0.53	0.34	0.42	0.22
SDXL	50	0.59	31.40	30.35	0.58	0.72	0.38	0.54	0.34	0.42	0.23
SD 3.5 large	10	0.51	31.42	37.56	0.71	0.90	0.46	0.63	0.27	0.34	0.19
SD 3.5 large	30	1.49	**31.90**	22.77	0.48	0.56	0.39	0.71	0.30	0.30	0.31
SD 3.5 large	50	2.45	31.69	22.48	0.55	0.58	0.53	0.71	0.41	0.34	0.47
FLUX.2 klein base 9b	10	0.82	31.71	22.93	0.45	0.55	0.31	0.79	0.19	0.21	0.18
FLUX.2 klein base 9b	30	2.69	31.81	**21.27**	0.56	0.56	0.56	0.81	0.42	0.30	0.51
FLUX.2 klein base 9b	50	4.55	31.78	21.57	0.78	0.66	0.88	0.82	0.68	0.45	0.85
FLUX.2 klein 4b	2	**0.05**	31.44	25.32	0.48	0.60	0.30	0.81	0.13	0.17	*0.09*
FLUX.2 klein 4b	6	0.14	31.22	24.01	0.45	0.57	0.29	0.81	0.13	0.17	*0.09*
FLUX.2 klein 4b	12	0.27	31.21	22.63	0.43	*0.54*	**0.27**	0.79	0.15	0.19	0.11
FLUX.2 klein 9b	2	*0.10*	*31.83*	23.14	0.44	0.55	*0.28*	**0.84**	**0.11**	**0.14**	**0.07**
FLUX.2 klein 9b	6	0.28	31.52	22.46	*0.42*	*0.54*	**0.27**	**0.84**	*0.12*	*0.15*	*0.09*
FLUX.2 klein 9b	12	0.54	31.50	*21.43*	**0.41**	**0.51**	**0.27**	*0.83*	0.14	0.16	0.13

Compositional benchmark with GenEval. Since FID compares closeness to a data distribution rather than a compositional benchmark of image quality, we also conduct an evaluation with GenEval (Ghosh et al., 2023). This benchmark evaluates whether generated images respect compositional prompts. We evaluate 60 configurations using the prompts from the original paper. We record GenEval's overall score and sub criteria such as *position*, *color attribution*, *single object*, *two objects*, *counting*, and *colors*, enabling us to evaluate how the quality-energy trade-off varies by semantic criteria. An efficient configuration for the criteria *single object* may not be efficient for *position* for instance (Fig. 3).

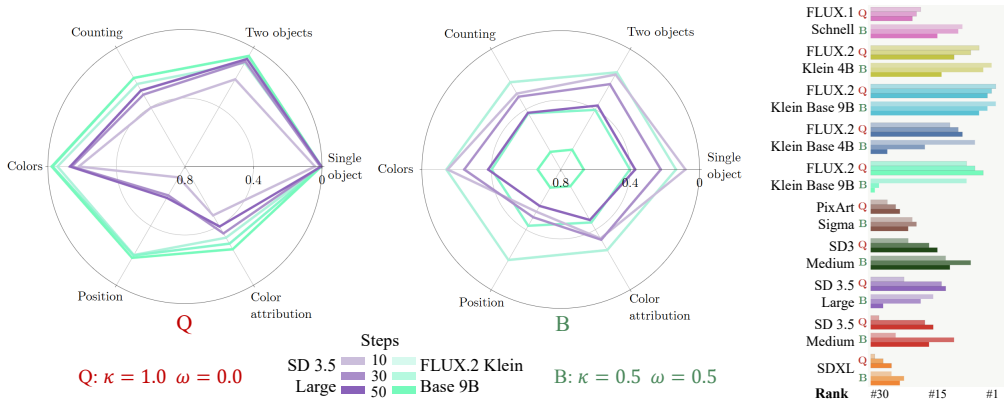


Figure 3: GenEval benchmark. **Left/Center**: Looking at quality only (Q, left), *FLUX.2 Klein Base 9B* is more robust to fewer steps than *SD 3.5 Large*, as GenEval curves are all close. In contrast, the balanced GQD (B, center) identifies interesting configurations for which quality is similar but the consumed energy is lower. **Right**: Rankings of the models in both settings ($Q|B$); in Balanced, configurations with many steps are ranked lower.

Tables 1 and 4 report the resulting values for both FID and GenEval as the quality metric of GQD, under three settings. The rankings illustrate how a quality-only metric can elect the worse models energy wise (*FLUX.2 klein base 9b* - 30 steps), while our GQD effectively selects models that perform well on both axes (*FLUX.2 klein 9b*). Figures 9 and 10 in Appendix further complement this insight. For the models robust to low denoising budgets, **a similar perceptual and compositional quality** (within 10%) is achieved with **a** 80% **drop in energy consumption** by drastically reducing the number of steps. Interestingly, we found that CLIPScore is not discriminative enough compared to FID and GenEval.

4.2 Video generation benchmark

Our video benchmark evaluates text-to-video (t2v) models with VBench on 7 models, 3 resolutions, 2 denoising steps. The measured energy ranges from 0.57 to 198.14 Wh per video. Energy-aware evaluation is therefore particularly important for video generation: output length, frames number, resolution, and temporal processing can amplify consumed energy far beyond t2i models. Resolution in particular can have a quadratic impact (Fig. 4).

VBench provides criteria such as visual quality, temporal consistency, motion smoothness, temporal flickering, and style. In Table 2, the GQD benchmark with VBench shows that for a slight quality gain, the energy consumed can be 133× higher (*Wan2.2 t2v A14B* 25 steps *vs LTX Video* 10 steps), highlighting the importance of the composite GQD metric. Figure 5 illustrates the denoising steps impact on each criterion for two models, and on the GQD ranking. Only some criteria improve with more steps, and our composite distance helps identifying efficient architectures among the good performers quality-wise.

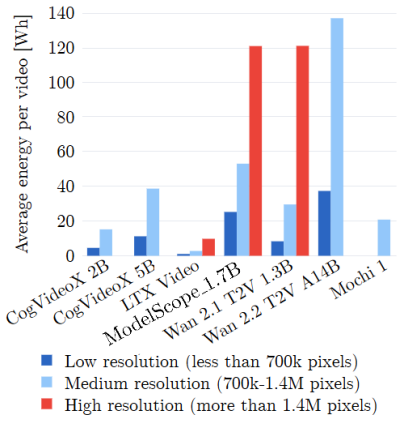


Figure 4: Consumed energy per generated video resolution. High resolutions lead to substantially higher energy consumption.

Table 2: Video benchmark at medium resolution. GQD (VBench) under three weightings: Balanced ($\kappa = \omega = 0.5$), Quality-first ($\kappa = 0.8,\ \omega = 0.2$), and Energy-first ($\kappa = 0.2,\ \omega = 0.8$). Color bars indicate the per-column ranking (larger is better).

Model	Steps	VBench ↑	Energy ↓	GQD B↓	GQD Q↓	GQD E↓
CogVideoX 2B	10	0.54	9.38	0.32	0.41	0.21
CogVideoX 2B	25	0.57	22.01	0.32	0.39	0.21
CogVideoX 5B	10	0.49	22.57	0.37	0.46	0.25
CogVideoX 5B	25	0.54	55.00	0.37	0.43	0.30
LTX Video	10	0.58	**1.49**	0.30	0.38	0.19
LTX Video	25	0.58	*3.46*	0.30	0.38	*0.19*
Mochi 1	16	0.54	12.14	0.33	0.41	0.21
Mochi 1	40	0.62	29.30	*0.29*	*0.35*	0.21
ModelScope t2v 1.7B	10	0.54	30.55	0.34	0.42	0.24
ModelScope t2v 1.7B	25	0.55	75.21	0.40	0.43	0.37
Wan2.1 t2v 1.3B	10	0.61	16.66	**0.28**	0.35	**0.19**
Wan2.1 t2v 1.3B	25	**0.63**	40.43	0.29	**0.34**	0.23
Wan2.2 t2v A14B	10	0.62	78.92	0.37	0.38	0.36
Wan2.2 t2v A14B	25	*0.63*	195.97	0.68	0.52	0.82

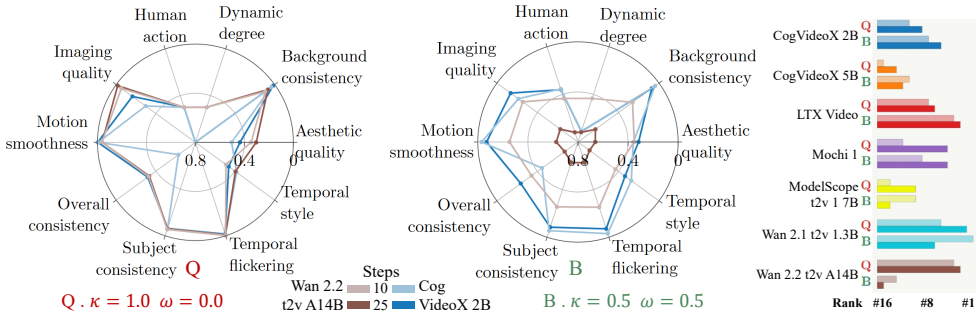


Figure 5: VBench benchmark. **Left/Center**: looking at quality only (Q, left), *CogVideoX 2B* is slightly behind *Wan2.2 A14B*, but the balanced setting (B, center) shows that it requires less energy and is thus ranked way better. **Right**: Rankings of all models in both settings (Q|B); in balanced, configurations with many steps are often ranked lower.

5 Extension to complete environmental cost

Power capping: energy-latency trade-off. We study GPU power capping as a configuration variable. For a fixed model, step count, and resolution, we sweep the GPU's maximum power draw from 100% down to 40% of its nominal value (400W for the A100 SXM4 80GB GPU used in our experiments). This parameter keeps the quality untouched,

but affects latency and energy. Lowering the power cap typically increases latency, since computations take longer. However, the resulting drop in power draw can, in some cases, outweigh this latency increase, yielding lower total energy.

As shown in Fig. 6 and 13, the sweet spot for minimized energy varies across models but averages approximately at 60% GPU power: 21% energy savings for 23% additional latency. At 40%, latency more than doubles (+111%) while energy savings fall to 10%. This overall trend holds across tested models. Additional metrics are reported in Tab.5 and 6.

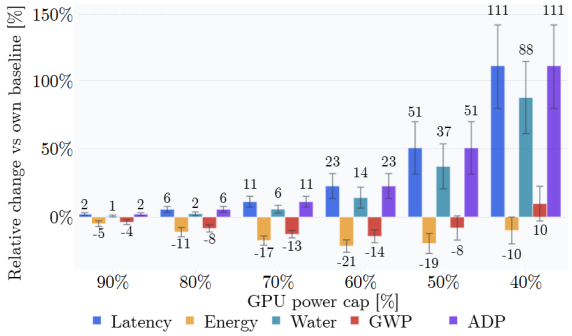


Figure 6: Relative impact of the power capping on the environmental impacts, compared to the baseline (100% power). After 60%, the energy gain decreases.

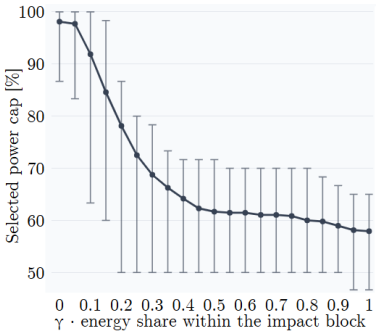


Figure 7: Power cap of the best configuration ranked by GQD_B depending on γ, averaged for all models.

Three-criteria score: quality, energy, and latency. Previously, latency was tied to energy since GPUs ran at full capacity. With power capping decoupling the two, it motivates a separate control over their relative weights, as they drive distinct environmental impacts: For instance, ADP is computed as the share of a GPU's lifetime consumed by a single inference, depending mostly on latency. We therefore introduce an *impact block* combining energy and latency, weighted by γ, that controls their relative share within GQD. As with energy, we normalize the latency gap $d_L(c)$ using a fixed budget L_{lim} as $d_L(c) = L(c)/L_{\text{lim}}$. The ideal three-criteria configuration is $(d_Q, d_E, d_L) = (0, 0, 0)$: **optimal quality**, **zero energy**, **zero latency**. We extend the ideal-distance score by defining:

$$GQD_{\kappa,\omega,\gamma}(c) = \sqrt{\kappa d_Q(c)^2 + \omega\left[\gamma d_E(c)^2 + (1-\gamma) d_L(c)^2\right]} \tag{6}$$

This formulation enables studying the impact of γ on the models of interest, and adjusting power capping depending on the objective (Fig. 7 and 14). This is particularly interesting in the context of jobs orchestration for favoring either latency or energy, and enables a further 20% energy reduction, that applies on the 80% already saved for some models, totaling 84%.

6 Adaptation to LLMs

Intelligence per Watt (or Joule) formulation. A formulation in the spirit of Saad-Falcon et al. (2026) could be used for images and videos, with a quality metric, but it poorly adapts to these domains. *Quality per Watt* would be easily biased by capping the GPU power, affecting the power term, and the resulting metric. *Quality per Joules* would make more sense, but Fig. 8 and 15 (Appendix C) show that it is fully biased towards promoting sober configurations, ignoring quality. This ratio may be more balanced for LLMs, which requires less energy, but is not informative for images and videos. Moreover, these formulations prevent controlling the quality-environmental trade-off, which GQD permits.

Adapting GQD to Language Models. To demonstrate that GQD can also serve LLMs evaluation, we use GreenServ (Ziller et al., 2026), which reports both quality and energy data for 16 open-access LLMs on 5 tasks, with 500 queries per task. We plug their task-level per-model accuracy and energy measures in Eq.5, and report rankings in Table 3: IPJ, energy, GQD, and accuracy. IPJ closely follows energy, suggesting a potential bias even for LLMs, while GQD balances energy and accuracy meaningfully.

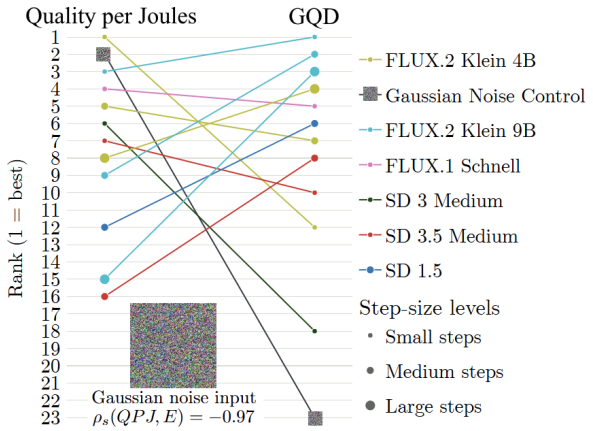


Figure 8: *Quality per Joules* (QPJ) versus $GQDB$ rankings: QPJ is fully biased towards energy, favoring fewer steps for all models, to the point of ranking 2^{nd} a Gaussian noise generator (the high correlation $\rho_s(QPJ, E)$ confirms this bias, see C). GQD ranks it last, balancing quality and energy.

Model	IPJ↓	energy↓	GQDB↓	accur.↓
Llama 3.2 1B	**1**	*2*	11	14
Qwen 2.5 0.5B	*2*	**1**	13	16
Qwen 2.5 1.5B	3	3	4	10
Phi-4 Mini 4B	4	5	**1**	5
Llama 3.2 3B	5	4	6	11
Qwen 2.5 3B	6	6	*2*	6
Mistral 7B	7	7	5	9
Llama 3.1 8B	8	11	3	4
Gemma 3 4B	9	9	7	7
Qwen 2.5 7B	10	10	8	8
Qwen 2.5 14B	11	13	9	3
Gemma 3 1B	12	8	12	12
Gemma 3 12B	13	14	10	*2*
Phi-4 14B	14	12	14	15
Gemma 3 27B	15	16	16	**1**
Yi 34B	16	15	15	13

Table 3: Rankings on GreenServ measures on 16 LLMs between IPJ, energy, GQDB, and accuracy. IPJ closely follows energy, suggesting a possible bias, while GQD balances energy and accuracy.

7 Limitations and Future Work

Closed models. Handling cloud-based proprietary models is challenging for two reasons: the ability to measure their energy per inference is limited, and they may not expose the ability to tweak the inference configuration. We could explore in future work how latency and credits that are charged to consumers may be used as proxies for energy consumption.

Incomplete Life Cycle Assessment. Even if our Green Quality Distance incorporates embodied and operational impacts, it is not yet a complete LCA of a model, leaving apart the training costs, the idle time and the complete infrastructure allowing to run a GPU. Future work could embed those as a share per inference.

8 Discussion and Conclusion

We present Green Quality Distance, a novel metric for evaluating configurations of image and video generative models simultaneously by the quality of the results they produce and the environmental impact of running the inference. Our normalized formulation as a distance to the ideal is flexible in terms of the particular quality metric or environmental cost to use, and can adapt to various domains and objectives. We present several benchmarks, first focusing on a quality-energy objective for images and videos, and further extend it to the triplet quality-energy-latency that allows including the common environmental indicators: Abiotic Depletion Potential, Global Warming Potential, and Water Use.

One key component of our approach is the explicit evaluation of model configurations instead of a single model instance. In this context, we explored the impact of hyperparameters on the quality-impact trade-off, and revealed the huge influence of the denoising budget, possibly exposed as a proxy for quality, without a clear insight inon the possible resource gain or waste. The outcome is clear: for the recent models tested (distilled or not), up to 81% of the energy may be saved while maintaining a similar perceptual quality. We also demonstrate that capping the GPU power can save up to 20% of energy at a similar loss in latency, allowing a job orchestrator to favor one or the other in an informed way.

Thereby, the Green Quality Distance has proved to be useful in identifying efficient configurations that exploit the full potential of generative image and video models while avoiding resource waste. We further analyze our formulation on LLMs and demonstrate that it is compatible with this field, as long as the quality metric is adapted for the task.

9 AI Use

In this work, we used generative AI tools for: design or provide feedback on research methodology or experiments, implement methods, assist with translation. We have not used generative AI tools for: help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, support qualitative and thematic data analysis, interpret results. Generate synthetic data sets, clean and reformat dataset are not applicable to this work. Additionally, we used generative AI tools for: summarize or analyze existing literature and identify relevant literature.

We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness, literature was mostly manually parsed and reviewed, and the use of AI for the manuscript was limited to rephrasing, polishing and spell checking. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

10 Reproducibility

We plan to release our source code on GitHub for measuring the energy consumption and aggregating the embodied impact, as described in Sec. 3, as well as for computing the Green Quality Distance given the different measures. We will also share the results of the benchmarks presented in section 4, including the code for downloading the subset of COCO2017 dataset that we use, and a sample code for running an entire benchmark. In the meantime, we provide all this content as an archive in supplementary material.

We provide in Appendix B.1 a statistical calibration of our composite formulation, as well as important elements of reproducibility: the limit values E_{lim}, Q_{lim} and L_{lim} that we use.

References

Adobe. Inaugural adobe creators' toolkit report: 86 percent of global creators use creative generative ai, see it boosting creator economy, 2025. URL https://news.adobe.com/news/2025/10/adobe-max-2025-creators-survey.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, volume 2025, pp. 102075–102121, 2025.

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

Benoit Courty, Victor Schmidt, Sasha Luccioni, Goyal-Kamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, Amine Saboni, Inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Hugues de Lavoreille, Niko Laskaris, Edoardo Abati, Douglas Blank, Ziyao Wang, Armin Catovic, Marc Alencon, Michal Stechly, Christian Bauer, Lucas Otávio N. de Araújo, JPW, and MinervaBooks. mlco2/codecarbon: v2.4.1, May 2024. URL https://doi.org/10.5281/zenodo.11171501.

Julien Delavande, Regis Pierrard, and Sasha Luccioni. Video killed the energy budget: Characterizing the latency and power regimes of open text-to-video models, 2025. URL https://arxiv.org/abs/2509.19222.

EcoLogits. Ecologits calculator, 2026. URL https://ecologits.ai/latest/. Accessed: 2026-09-03.

Cooper Elsworth, Keguo Huang, David Patterson, Ian Schneider, Robert Sedivy, Savannah Goodman, Ben Townsend, Parthasarathy Ranganathan, Jeff Dean, Amin Vahdat, Ben

Gomes, and James Manyika. Measuring the environmental impact of delivering AI at google scale, 2025. URL https://arxiv.org/abs/2508.15734.

Sophia Falk, David Ekchajzer, Thibault Pirson, Etienne Lees-Perasso, Augustin Wattiez, Lisa Biber-Freudenberger, Sasha Luccioni, and Aimee van Wynsberghe. More than carbon: Cradle-to-grave environmental impacts of genai training on the nvidia a100 gpu. *Environmental Impact Assessment Review*, 121:108525, 2026.

Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the Content Bias in Fréchet Video Distance . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7277–7288. IEEE Computer Society, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: an object-focused framework for evaluating text-to-image alignment. NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

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. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2024. doi: 10.1109/CVPR52733.2024.02060.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Etienne Lees-Perasso, David Ekchajzer, Gauthier Roussilhe, and Thomas De Latour. Analyses de cycle de vie de gpu pour l'intelligence artificielle, 2026. URL https://data.hubblo.org/gpu.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Yan Liu and He Wang. Who on earth is using generative ai? *World Development*, 199: 107260, 03 2026. doi: 10.1016/j.worlddev.2025.107260.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the carbon footprint of BLOOM, a 176B parameter language model, 2022. URL https://arxiv.org/abs/2211.02001. Published in Journal of Machine Learning Research 24 (2023).

Sasha Luccioni, Yacine Jernite, and Emma Strubell. Power hungry processing: Watts driving the cost of ai deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 85–99, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658542. URL https://doi.org/10.1145/3630106.3658542.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? pp. 948–958, 2026.

NVIDIA Corporation. Nvidia management library (nvml). https://docs.nvidia.com/deploy/nvml-api/, 2026. NVIDIA GPU Deployment and Management Documentation.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Jon Saad-Falcon, Avanika Narayan, Hakki Orhun Akengin, J. Wes Griffin, Herumb Shandilya, Adrian Gamarra Lafuente, Medhya Goel, Rebecca Joseph, Shlok Natarajan, Etash Kumar Guha, Shang Zhu, Ben Athiwaratkun, John Hennessy, Azalia Mirhoseini, and Christopher Ré. Intelligence per watt: Measuring intelligence efficiency of local ai, 2026. URL https://arxiv.org/abs/2511.07885.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 2234–2242, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green ai. *Commun. ACM*, 63(12):54–63, November 2020. ISSN 0001-0782. doi: 10.1145/3381831. URL https://doi.org/10.1145/3381831.

Thibault Simon, David Ekchajzer, Adrien Berthelot, Eric Fourboul, Samuel Rincé, and Romain Rouvoy. BoaviztAPI: A bottom-up model to assess the environmental impacts of cloud services. In *HotCarbon'24 – 3rd Workshop on Sustainable Computer Systems*, Santa Cruz, CA, USA, 2024. URL https://hal.science/hal-04621947v2. HAL: hal-04621947.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://aclanthology.org/P19-1355/.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2019. URL https://arxiv.org/abs/1812.01717.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Thomas Ziller, Shashikant Ilager, Alessandro Tundo, Ezio Bartocci, Leonardo Mariani, and Ivona Brandic. Greenserv: Energy-efficient context-aware dynamic routing for multi-model llm inference, 2026. URL https://arxiv.org/abs/2601.17551.

A Benchmarks

In the following sections, we provide a more complete version of the benchmarks presented in the main text, as well as additional results regarding power capping and its impact on energy consumption and latency. The complete benchmarks are available in the supplementary material.

A.1 Image benchmarks

Table 4 shows our image benchmark in 1024^2 pixels, with CLIPScore, FID, and GenEval metrics, for three number of inference steps.

The benchmark in 512^2 includes three more models, unavailable in 1024^2: *SD 1.5*, *PixArt-Sigma* and *FLUX Schnell*. The full table of results is not reported here but presented similar outcome than in 1024^2. Fig. 9 is an extract of the 10 best performing configurations in 512^2 according to GQD computed with FID metric, with two quality-energy tradeoff settings. Fig. 10 shows the distance to the ideal configuration for image generation tasks in 512^2 with the GenEval score as the quality metric.

Table 4: Extended version of Tab. 1. Image benchmark over 8 models with 3 numbers of inference steps, at 1024^2 pixels. Bars indicate the relative ranking within each column, with longer bars corresponding to better configurations. We report three quality metrics (FID, CLIPScore, GenEval), energy consumption, and our GQD for both metric, under three weightings: Balanced ($\kappa = \omega = 0.5$), Quality-first ($\kappa = 0.8$, $\omega = 0.2$), and Energy-first ($\kappa = 0.2$, $\omega = 0.8$).

Model	step #	energy ↓	CLIP Score↑	FID ↓	GQD B↓	GQD Q↓	GQD E↓	Gen Eval↑	GQD B↓	GQD Q↓	GQD E↓
SDXL	10	0.13	31.44	28.60	0.54	0.68	0.34	0.48	0.37	0.46	0.23
SDXL	30	0.36	31.44	29.33	0.55	0.70	0.36	0.53	0.34	0.42	0.22
SDXL	50	0.59	31.40	30.35	0.58	0.72	0.38	0.54	0.34	0.42	0.23
SD 3 medium	10	0.16	31.57	27.40	0.52	0.65	0.33	0.63	0.26	0.33	0.17
SD 3 medium	30	0.46	31.76	26.98	0.51	0.64	0.33	0.69	0.23	0.28	0.16
SD 3 medium	50	0.76	31.72	25.83	0.50	0.62	0.34	0.69	0.24	0.28	0.20
SD 3.5 medium	10	0.23	31.34	31.25	0.59	0.74	0.37	0.50	0.36	0.45	0.23
SD 3.5 medium	30	0.66	31.86	24.26	0.47	0.58	0.31	0.68	0.24	0.29	0.19
SD 3.5 medium	50	1.10	31.83	24.92	0.50	0.60	0.36	0.69	0.27	0.29	0.25
SD 3.5 large	10	0.51	31.42	37.56	0.71	0.90	0.46	0.63	0.27	0.34	0.19
SD 3.5 large	30	1.49	**31.90**	22.77	0.48	0.56	0.39	0.71	0.30	0.30	0.31
SD 3.5 large	50	2.45	31.69	22.48	0.55	0.58	0.53	0.71	0.41	0.34	0.47
FLUX.2 klein base 4b	10	0.46	31.12	28.71	0.54	0.69	0.35	0.76	0.18	0.22	0.14
FLUX.2 klein base 4b	30	1.39	31.50	27.21	0.55	0.66	0.41	0.77	0.26	0.24	0.28
FLUX.2 klein base 4b	50	2.28	31.51	27.55	0.62	0.69	0.53	0.78	0.37	0.29	0.43
FLUX.2 klein base 9b	10	0.82	31.71	22.93	0.45	0.55	0.31	0.79	0.19	0.21	0.18
FLUX.2 klein base 9b	30	2.69	31.81	**21.27**	0.56	0.56	0.56	0.81	0.42	0.30	0.51
FLUX.2 klein base 9b	50	4.55	31.78	21.57	0.78	0.66	0.88	0.82	0.68	0.45	0.85
FLUX.2 klein 4b	2	**0.05**	31.44	25.32	0.48	0.60	0.30	0.81	0.13	0.17	*0.09*
FLUX.2 klein 4b	6	0.14	31.22	24.01	0.45	0.57	0.29	0.81	0.13	0.17	*0.09*
FLUX.2 klein 4b	12	0.27	31.21	22.63	0.43	*0.54*	**0.27**	0.79	0.15	0.19	0.11
FLUX.2 klein 9b	2	*0.10*	31.83	23.14	0.44	0.55	*0.28*	**0.84**	**0.11**	**0.14**	**0.07**
FLUX.2 klein 9b	6	0.28	31.52	22.46	*0.42*	*0.54*	**0.27**	**0.84**	*0.12*	*0.15*	*0.09*
FLUX.2 klein 9b	12	0.54	31.50	*21.43*	**0.41**	**0.51**	**0.27**	*0.83*	0.14	0.16	0.13

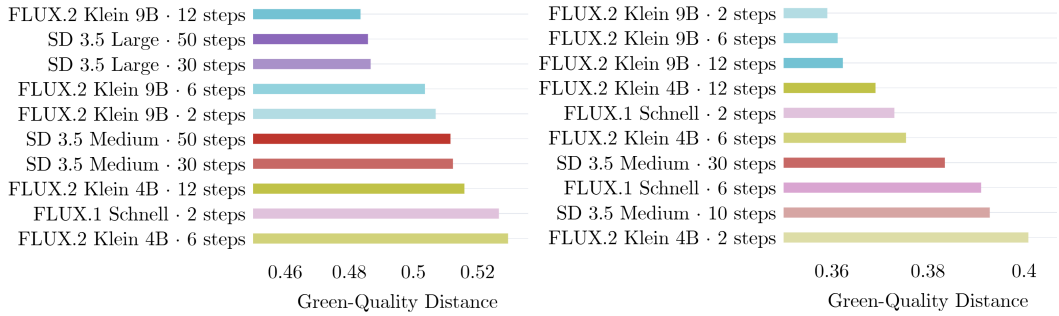


Figure 9: Top 10 configurations of Green Quality Distance computed with FID for image generation at 512^2 pixels. Left: pure quality $\kappa = 1$, $\omega = 0$. Right: balanced trade-off $\kappa = \omega = 0.5$. Even if the winner is *FLUX.2 klein 9B* for both, the winning configuration differs. Also, *StableDiffusion 3.5 Large*, ranked second in the quality benchmark, is not even in the balanced GQD top 10.

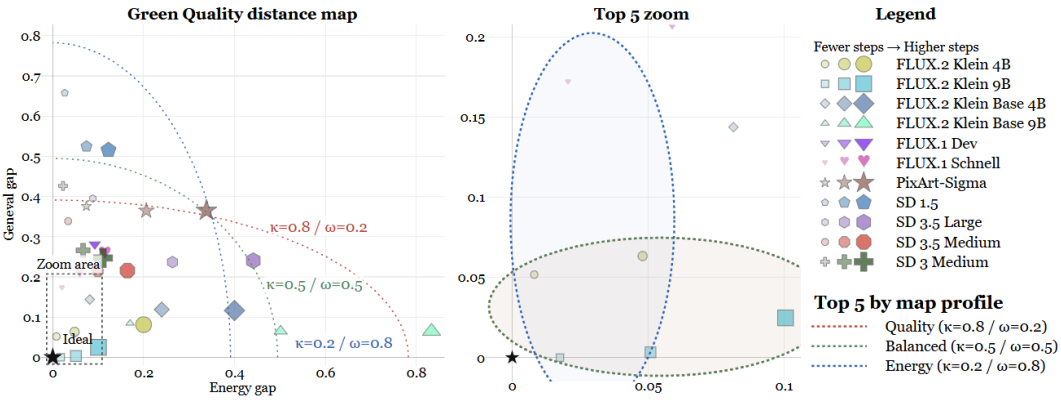


Figure 10: Green Quality Distance: Distance between benchmarked configurations and ideal configuration $\star$ for image generation at 512^2 resolution, using GenEval score for quality. Ellipses are iso-distances weighted by κ and ω, showing how the distance to $\star$ and the Top 5 evolve for different parameters.

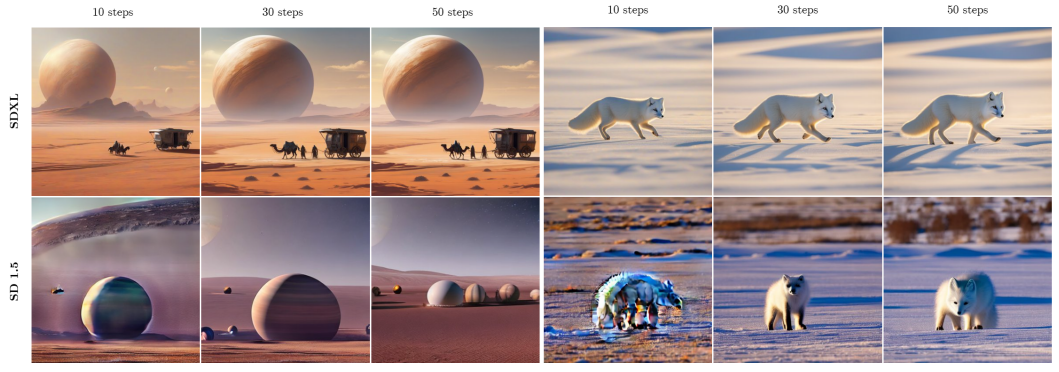


Figure 11: Qualitative comparison of the denoising budget impact on image generations. Prompts are: "a desert caravan crossing a salt plain beneath a giant ringed planet, cinematic concept art" and "an arctic fox trotting across a snow plain in low winter sun, wildlife telephoto photo".

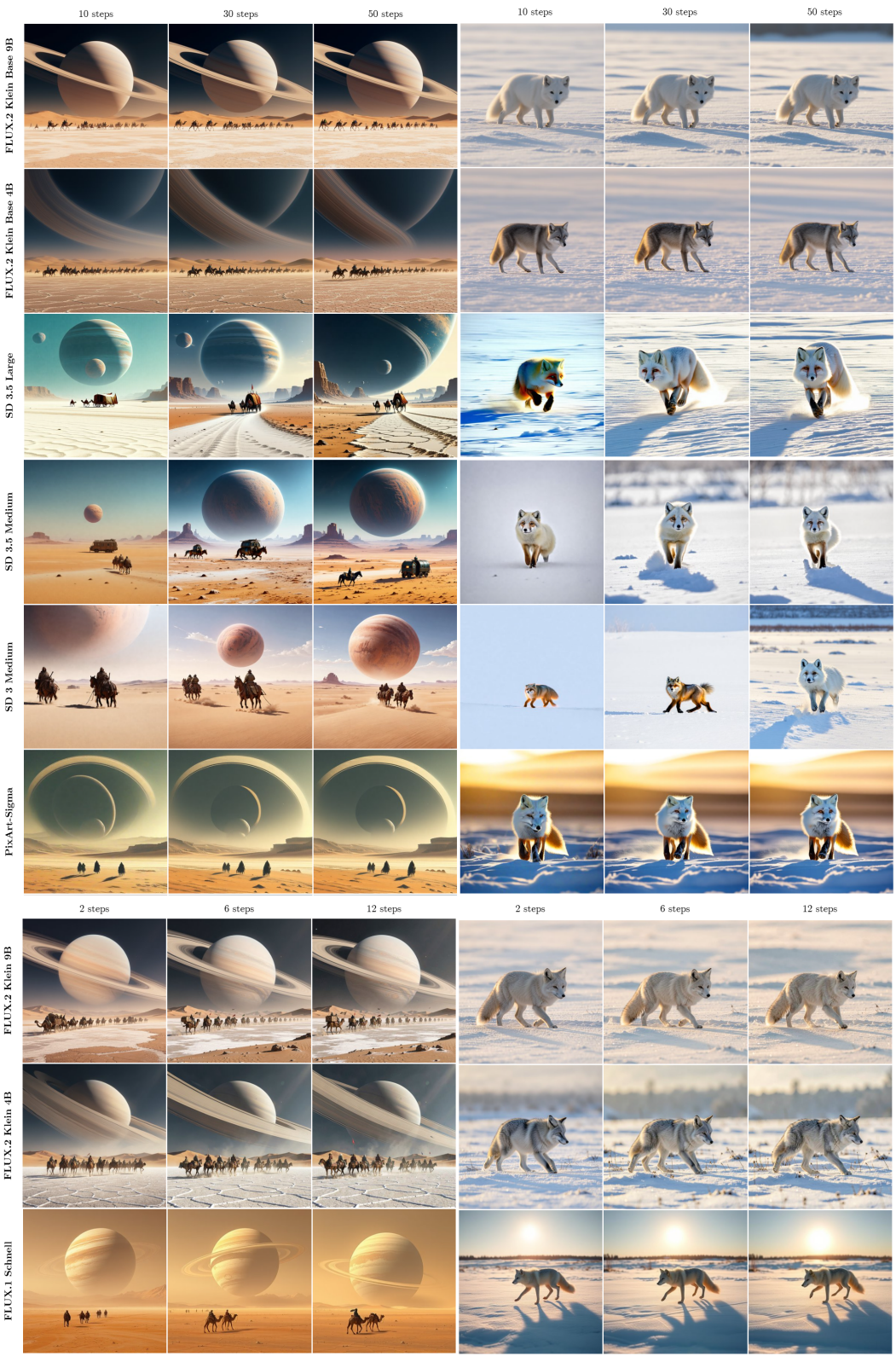


Figure 12: Qualitative comparison of the denoising budget impact on image generations. Prompts are: "a desert caravan crossing a salt plain beneath a giant ringed planet, cinematic concept art" and "an arctic fox trotting across a snow plain in low winter sun, wildlife telephoto photo".

A.2 Power capping benchmark

Figure 13 reveals the energy-latency frontier of three different configurationsTable 5 then compares the energy-optimal power caps across models and their effects on GWP, latency, ADPe, and water use, while Table 6 reports the corresponding absolute impacts across all tested caps for two configurations. Finally, Figure 14 illustrates how the energy-latency weighting parameter γ shifts the preferred configurations with the GQD.

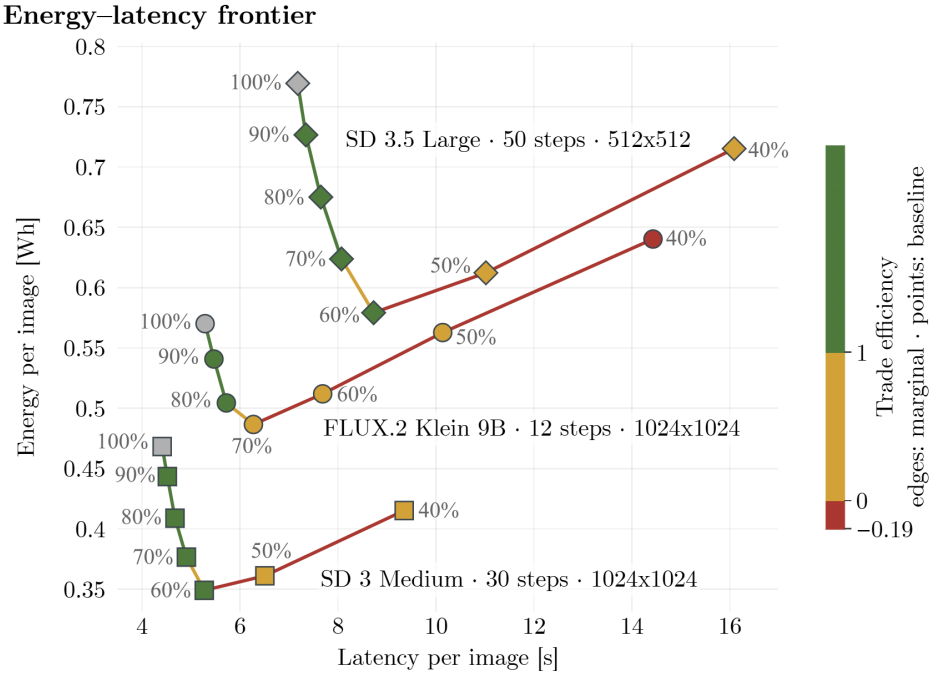


Figure 13: Influence of power capping on energy consumption and latency for three configurations. For all, reducing the power limit decreases the energy required per generation at the cost of increased latency. The color of points represents the relative change in energy compared to the baseline (100% GPU, the gray point) The color of segments represents the relative change in energy consumption with respect to the relative change in latency, highlighting configurations for which power capping provides a more favorable energy–latency trade-off.

Table 5: Per-model trade-off at the energy-optimal power cap, relative to the uncapped 100% baseline. Lowering the cap cuts energy and its carbon footprint (GWP) at the cost of higher latency and water use; ADPe scales exactly with latency, so both share a single column. Impact factors use the us-west-2 (Oregon) grid at 0.351 $kgCO_2eq/kWh$, a water-usage effectiveness of 1.15 L/kWh, and a 4-year GPU lifetime (Elsworth et al., 2025).

Model	Opt. cap [%]	Energy [Wh]	Energy Δ%	GWP Δ%	Lat./ADPe Δ%	WU Δ%
SD 1.5	50	0.095	**−32.3**	−31.0	+5.0	−1.1
SDXL	60	0.406	−27.8	−26.6	+18.5	+9.3
SD 3 Med	60	0.349	−25.5	−24.2	+19.5	+10.7
SD 3.5 Med	60	0.492	−27.4	−26.1	+18.7	+9.5
SD 3.5 Large	60	1.234	−19.7	−18.3	+29.3	+19.6
FLUX Schnell	60	0.252	−19.7	−18.3	+30.4	+20.5
FLUX.2 Klein 4B	70	0.237	−18.7	−17.8	+13.6	+7.2
FLUX.2 Klein 9B	70	0.487	−14.7	−13.8	+18.7	+12.1

Table 6: Effect of lowering the GPU power cap, per generated image. Percentages (in gray) are relative to the uncapped 100% setting. Latency rises steadily as the cap decreases, while energy and its derived impacts reach a minimum near 60% before climbing again. ADPe scales exactly with latency. Impact factors use the us-west-2 (Oregon) grid at 0.351 $kgCO_2eq/kWh$, a water-usage effectiveness of 1.15 L/kWh, and a 4-year GPU lifetime (Elsworth et al., 2025).

Model	Cap [%]	Latency [s]	Energy [Wh]	GWP [$mgCO_2eq$]	ADPe [ng Sb-eq]	WU [mL]
SD 3.5 L 30 steps	100	14.33	1.536 (+0.0%)	554.2 (+0.0%)	916 (+0.0%)	8.96 (+0.0%)
	90	14.76	1.452 (−5.4%)	525.3 (−5.2%)	944 (+3.0%)	9.07 (+1.3%)
	80	15.43	1.362 (−11.3%)	494.4 (−10.8%)	987 (+7.7%)	9.31 (+4.0%)
	70	16.35	1.264 (−17.7%)	460.9 (−16.8%)	1045 (+14.1%)	9.66 (+7.8%)
	60	18.52	**1.234** (−19.7%)	**452.6** (−18.3%)	1184 (+29.3%)	10.71 (+19.6%)
	50	23.31	1.291 (−16.0%)	477.6 (−13.8%)	1491 (+62.7%)	13.18 (+47.2%)
	40	32.96	1.463 (−4.8%)	548.1 (−1.1%)	2108 (+130.1%)	18.22 (+103.5%)
FLUX.2 Klein 9B 12 steps	100	5.29	0.570 (+0.0%)	205.7 (+0.0%)	338 (+0.0%)	3.31 (+0.0%)
	90	5.47	0.541 (−5.1%)	195.6 (−4.9%)	350 (+3.4%)	3.37 (+1.7%)
	80	5.72	0.504 (−11.5%)	183.1 (−11.0%)	366 (+8.2%)	3.45 (+4.3%)
	70	6.27	**0.487** (−14.7%)	**177.4** (−13.8%)	401 (+18.7%)	3.71 (+12.1%)
	60	7.69	0.512 (−10.2%)	187.8 (−8.7%)	492 (+45.4%)	4.45 (+34.4%)
	50	10.14	0.563 (−1.3%)	208.2 (+1.2%)	648 (+91.8%)	5.73 (+73.3%)
	40	14.43	0.640 (+12.3%)	239.9 (+16.7%)	923 (+173.0%)	7.98 (+141.1%)

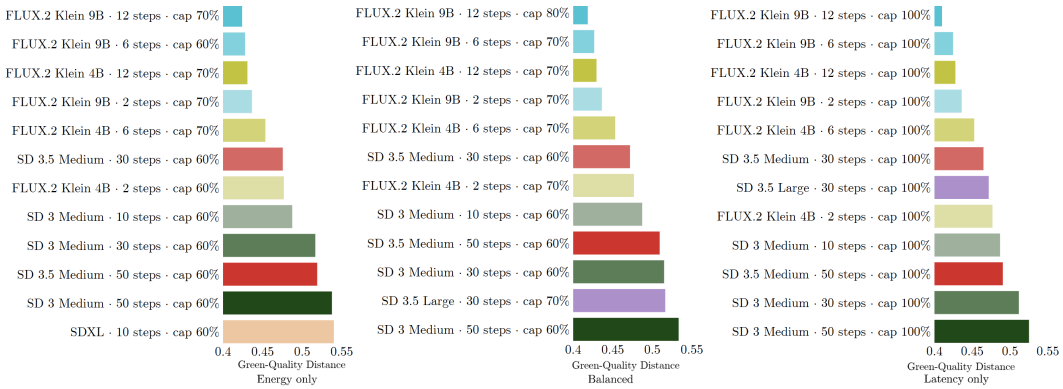


Figure 14: Impact of the sharing parameter γ on the ranking, keeping only the best configuration of each model. From left to right, γ decreases from 1 (energy only) through 0.5 (balanced) to 0 (latency only). As γ decreases, the preferred power cap shifts from the energy-optimal region (60 to 70%) towards 100%, the lowest-latency and ADP-minimizing operating point. In the three columns, $\kappa = 0.5$, $\omega = 0.5$.

B Green Quality Distance formulation

B.1 Statistical calibration and reproducibility

For the two components, quality and energy, of *GQD* to contribute comparably, they must be rescaled onto a common scale. We avoid standard-score normalization (mean, std), which would make scores population-dependent and thus incomparable across studies. We instead rescale them against fixed reference limits (Q_{lim}, E_{lim}), that we calibrate once and share. They act as reference scales rather than strict bounds: distances need not lie in [0,1], and values above one indicate a configuration exceeding its reference limit.

We assess whether energy or quality dominates using a Spearman rank correlation, defined for two variables X and Y as the Pearson correlation between their rank R:

$$\rho_s(X,Y) = \mathrm{Corr}(R[X], R[Y]) = \frac{\mathrm{Cov(R[X],R[Y])}}{\sigma_{R[X]}\sigma_{R[Y]}} \tag{7}$$

We compute the correlation of GQD with each criterion: $\rho_Q = \rho_s(GQD_{\kappa,\omega}, d_Q)$, $\rho_E = \rho_s(GQD_{\kappa,\omega}, d_E)$, a high absolute value for the correlation indicates association between both criteria. We define $\Delta_\rho = |\rho_Q - \rho_E|$ as a diagnostic of trade-off symmetry, with $\Delta_\rho = 0$ indicating a perfectly balanced contribution.

Rather than setting the reference limits arbitrarily, we calibrate them using a grid search: for each quality metric and resolution, E_{lim} and Q_{lim} are chosen so as to minimize Δ_ρ under the symmetric weighting $\kappa = \omega = 0.5$. For GenEval at 1024^2 resolution we obtain $\rho_Q = 0.54$, $\rho_E = 0.53$, and $\Delta_\rho \approx 0$. This near-zero gap indicates that for this benchmark, GQD is comparably associated with quality and energy efficiency.

Although this grid-search calibration avoids an explicit dependence on the mean and standard deviation of a particular sample, it still depends on the set of configurations used for calibration. Therefore, for strict cross-study comparability, we strongly recommend reusing the fixed reference values rather than recalibrating them for each new benchmark. We provide these values in Tab. 7 for all the experimented tasks.

Table 7: Reference limits used for the different GQD benchmarks. E_{lim} is expressed per generated output. L_{lim} is only used for the three-criterion power-capping benchmark.

Benchmark	Resolution	Q_{lim}	E_{lim} (Wh/output)	L_{lim} (s/output)
FID	512^2	48.4477	0.8252	–
	1024^2	37.5562	4.8262	–
KID	512^2	0.020016	0.6876	–
	1024^2	0.013333	3.3693	–
GenEval	512^2	1.0000	0.6112	–
	1024^2	1.0000	4.8262	–
VBench	Low	1.0000	65.3128	–
	Medium	1.0000	219.4825	–
	High	1.0000	344.7866	–
Power capping (FID)	512^2	48.4477	0.8252	10.5525
	1024^2	37.5562	4.8262	23.8102
GreenServ (LLM)	–	1.0000	0.3283	–

B.2 Theoretical properties of Green Quality Distance

Here we provide a compact characterization of the ranking induced by GQD. We assume $\kappa, \omega > 0$ and use the normalized gaps $d_Q(c)$ and $d_E(c)$ defined in Eq.4 of Section 3.3.

Effective weighting parameter. Since the square root is strictly increasing, ranking configurations by GQD is equivalent to ranking them by:

$$S_\lambda(c) = \lambda d_Q(c)^2 + d_E(c)^2, \qquad \lambda = \frac{\kappa}{\omega} \tag{8}$$

Hence, for fixed normalization references, only the relative weight λ matters for ranking. For minimized homogeneous metrics, such as FID or KID, using $d_Q = Q/Q_{\mathrm{lim}}$ and $d_E = E/E_{\mathrm{lim}}$ gives the equivalent raw-scale score $S_\theta(c)$:

$$S_\theta(c) = \theta Q(c)^2 + E(c)^2, \qquad \theta = \frac{\kappa E_{\mathrm{lim}}^2}{\omega Q_{\mathrm{lim}}^2} \tag{9}$$

Thus, weights and normalization limits jointly determine a single effective parameter θ.

Pairwise rank switches. Consider two non-dominated configurations c_1, c_2 such that $d_Q(c_1) < d_Q(c_2)$ and $d_E(c_1) > d_E(c_2)$. Their ranking changes at the unique threshold:

$$\lambda^\star = \frac{d_E(c_1)^2 - d_E(c_2)^2}{d_Q(c_2)^2 - d_Q(c_1)^2} \tag{10}$$

For $\lambda < \lambda^\star$, the lower-energy configuration is preferred; for $\lambda > \lambda^\star$, the higher-quality one is. Since the benchmark contains a finite number of configurations, these pairwise thresholds partition the weighting space into intervals over which the complete ranking is constant.

Pareto robustness. If c_1 and c_2 satisfy $d_Q(c_1) \leq d_Q(c_2)$ and $d_E(c_1) \leq d_E(c_2)$, with at least one strict inequality, then $S_\lambda(c_1) < S_\lambda(c_2), \forall \lambda > 0$. Therefore, a Pareto-dominated configuration (here, c_2) can never be optimal under GQD for strictly positive κ and ω.

Maximized quality metrics. The results above directly apply to maximized metrics (GenEval, VBench) through the gap $d_Q = 1 - Q/Q_{\text{lim}}$ defined in Eq.4. For a fixed Q_{lim}, the ranking therefore still depends only on λ, and the same switch-threshold and Pareto properties hold. Unlike the minimized homogeneous case, however, changing Q_{lim} cannot generally be absorbed into λ, since:

$$\left(1 - \frac{Q}{Q_{\text{lim}}}\right)^2 = 1 - \frac{2Q}{Q_{\text{lim}}} + \frac{Q^2}{Q_{\text{lim}}^2}. \tag{11}$$

The quality reference Q_{lim} must therefore remain explicit for maximized metrics.

C Comparison with Quality per Joules

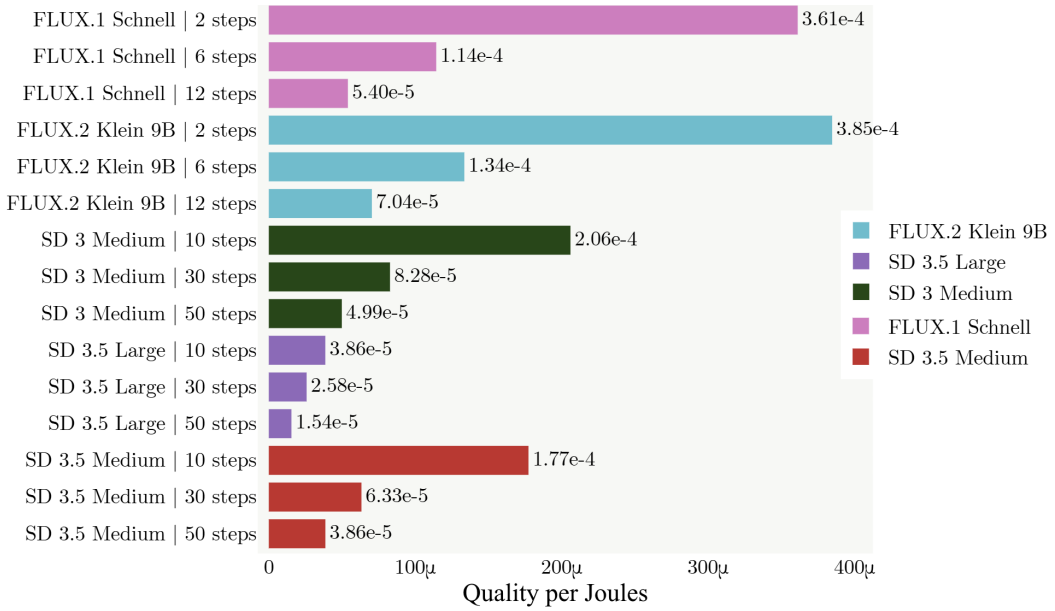


Figure 15: Evolution of Quality per Joules over different numbers of steps. The configuration with the fewest diffusion steps obtain consistently a higher Quality per Joules, revealing that this metric has a strong bias toward energy, possibly ignoring the resulting quality.

Computing the Spearman rank correlation of Quality per Joules with its two components, quality and energy, demonstrates that it is strongly correlated with the energy axis, with $\rho_s(\text{QPJ}, E) = -0.97$ whereas its correlation with quality, measured with FID here, is negligible: $\rho_s(\text{QPJ}, Q) = 0.108$. The negative sign for the energy correlation is expected because QPJ is maximized while energy wants to be minimized. Its large absolute absolute value therefore indicates a strong ranking agreement. These results confirm that QPJ is almost entirely driven by energy consumption rather than output quality.