GENO: Diversity-Preserving Green Optimization for Multi-Objective AI Service Orchestration

Anonymous authors
Paper under double-blind review

Abstract

The proliferation of AI inference services across edge and cloud infrastructures has elevated energy consumption and carbon emissions to critical concerns in service orchestration. On the AI-intensive five-objective formulations we study, conventional decomposition-based multi-objective optimizers exhibit what we term low-energy diversity collapse: neighborhood-restricted updates systematically under-explore the energy-efficient region of the Pareto front spanning inference energy, transport energy, carbon, latency, and reliability. To address this, we propose GENO, a green-aware service orchestration framework built around Decomposition–Dominance Coordination (DDC), a mechanism that synchronizes a dominance archive with the decomposition population via energy-priority redundancy removal and re-association. DDC is complemented by an energy-augmented state encoder, a five-objective formulation that partitions total energy into AI-inference and non-AI components, and a bounded weight-perturbation interface instantiated through a local language model. We establish two formal guarantees: DDC monotonically improves a green-weighted archive score (Proposition 1), and any preference-channel perturbation is Lipschitz-bounded in its effect on the scalarized objective (Proposition 2). Experiments on three service orchestration benchmarks show that, among the eight methods evaluated, GENO is the only one to reach a low-energy operating point (energy-per-request below 0.6 Wh and tail latency under 22 ms) on the AI-EO benchmark, improving energy-per-request over the strongest baseline by roughly a quarter to a third. Component ablation attributes -21.1% HV to DDC removal versus only -2.8% to the language interface, confirming that the gains are algorithmic.

1 Introduction

AI inference has become a dominant and persistent workload in modern service platforms, spanning edge video analytics, mobile intelligence, and cloud-hosted language services. Orchestration decisions now determine not only quality-of-service metrics such as delay and throughput, but also the energy consumed by model execution and the carbon intensity of the hosting infrastructure. Inference cost, once negligible relative to training, can become a significant contributor to both operational expenditure and environmental footprint at deployment scale (Desislavov et al., 2023; Luccioni et al., 2023). Concurrently, the electricity demand of data centers and 5G/6G networks continues to grow (Wu et al., 2022; Buzzi et al., 2016), while providers face increasing pressure to reconcile performance targets with decarbonization commitments (López-Pérez et al., 2022). Green service orchestration is therefore a systems-level challenge that cannot be reduced to single-device optimization.

Despite this need, three gaps persist. First, most orchestration methods optimize primarily for latency, throughput, or cost, and either treat energy and carbon as secondary metrics or ignore them altogether (Zhang & Li, 2007; Mao et al., 2016). Second, green computing studies typically address only one aspect of the problem—server power, offloading energy, or carbon-aware placement—and therefore cannot capture the coupled trade-offs among inference energy, transport energy, carbon emissions, service delay, and reliability (Liang

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.

et al., 2024; Qureshi et al., 2009; Zhang et al., 2016; Guo et al., 2018). Third, even when multi-objective search is employed, mainstream decomposition-based optimizers carry no explicit bias toward energy-efficient operating regions, causing them to under-explore solutions favored by hardware efficiency and time-varying carbon signals in AI-intensive workloads.

This paper introduces GENO, designed around one organizing principle: if green service orchestration is inherently a coupled many-objective problem, the search procedure should be structurally biased toward energy-efficient regions across modeling, representation, and optimization. GENO formulates orchestration as a five-objective problem over inference energy, network energy, carbon emissions, latency, and reliability. The same green signals that define the objectives are injected into the search through structured features and a decomposition–dominance coordination mechanism, encouraging the optimizer to preserve diverse but operationally meaningful low-energy solutions rather than pursuing purely geometry-driven Pareto coverage. We use search bias to denote any algorithmic mechanism that, with the objective vector fixed, alters the empirical visitation distribution of the search over feasible solutions toward a policy-relevant subregion (here, low-energy and low-carbon regions).

Under this formulation, language guidance plays a limited and optional role. Operators may hold high-level preferences that are hard to encode as fixed coefficients—for instance, temporarily favoring low-carbon nodes during high renewable availability. GENO exposes an auxiliary policy-guidance channel that perturbs objective weights within bounded ranges using structured directives from a local language model. The framework does not depend on such guidance, and the language model is not a standalone controller; it is a low-frequency mechanism for injecting policy preference when available, while the core optimization remains algorithmic.

Our main contributions, ranked by impact on the optimization outcome rather than narrative order, are:

1. Identification of low-energy diversity collapse. We isolate a structural pathology of MOEA/D-family optimizers on AI-intensive orchestration: under a five-objective green formulation, neighborhood-restricted updates concentrate the population in the latency-dominated region and leave the low-inference-energy Pareto front systematically unexplored, formalized via a subproblem alignment error (Eq. equation 7). We establish this empirically on the evaluated benchmarks and do not claim it as a universal property of all decomposition-based optimizers.

2. Decomposition–Dominance Coordination (DDC). The core novelty synchronizes a dominance archive with the decomposition population through (i) energy-priority Redundancy Removal that retains archive diversity along the inference-energy and carbon axes, and (ii) Energy-Priority Re-Association that rebinds subproblems to the current best-in-archive at each invocation epoch. Ablation attributes -21.1% HV and $+157.7\%$ EPR to DDC removal—the largest single-component effect.

3. Bounded weight-perturbation interface with a green-half-simplex bounded perturbation channel whose effect on the scalarized objective is provably Lipschitz (Proposition 2); a local LLM is one instantiation, contributing only -2.8% HV in ablation, confirming that GENO's gains are algorithmic.

4. Five-objective formulation in which $\mathcal{O}_1$ and $\mathcal{O}_2$ additively partition total system energy into inference and non-inference components without double counting; together with carbon, latency, and reliability, this captures coupled trade-offs single- or bi-objective methods cannot express.

5. Component-level evaluation. Against seven baselines on three benchmarks, GENO is the only evaluated method to reach the (EPR ≤ 0.6 Wh/req) $\wedge$ (tail latency $\leq$ 22 ms) Pareto corner on AI-EO, a roughly one-quarter-to-one-third EPR improvement over the strongest baseline.

2 Related Work

Energy footprint of AI inference. The environmental footprint of AI has spurred a "Green AI" movement (Schwartz et al., 2020). Strubell et al. (2019) quantified the carbon cost of NLP training, and subsequent work showed inference energy can dominate over the model lifecycle (Desislavov et al., 2023; Luccioni et al., 2023). Patterson et al. (2021) identified hardware efficiency (energy per FLOP), datacenter PUE, and grid carbon intensity as the three dominant levers, and Wu et al. (2022) argued that co-design of model, hardware, and placement is necessary. GENO operationalizes this co-design at the orchestration layer, jointly optimizing placement decisions affecting all three levers.

Carbon-aware placement and energy-efficient orchestration. Energy-efficient VNF placement has been studied via DVFS (Liang et al., 2024), server consolidation (Khosravi et al., 2013), and renewable-aware scheduling (Radovanović et al., 2023); MEC offloading (Zhang et al., 2016; Guo et al., 2018; Mao et al., 2017) targets device-side rather than inference-side energy, and geo-distributed placement (Qureshi et al., 2009) lacks multi-objective Pareto awareness. Grid carbon intensity varies up to 10× across regions (Lacoste et al., 2019; Lannelongue et al., 2021), motivating real-time carbon integration. Carbon-aware scheduling (Wiesner et al., 2021) and energy-efficient network design (López-Pérez et al., 2022) typically use single/bi-objective formulations incompatible with five-objective Pareto evaluation; GENO unifies these directions.

Multi-objective optimization and language-guided control. MOEA/D (Zhang & Li, 2007) and its variants (NSGA-III (Deb & Jain, 2014), MOEA/D-AGR (Qi et al., 2014)) exhibit recurrent failure modes in high-dimensional green settings (Ishibuchi et al., 2008): diversity collapse, solution–subproblem mismatch, and context-agnostic operators. Recent LLM-for-networks studies explore language-guided control (Mao et al., 2021) but do not separate settings where policy guidance is useful from those where it should remain auxiliary. GENO combines a five-objective formulation with a unified green search bias and bounded, optional policy guidance.

3 System Model and Problem Definition

We formalize multi-objective green service orchestration with awareness of AI inference workloads; key notation is summarized in Table 5 (Appendix A).

Service infrastructure and energy model. The substrate network $\mathcal{G}_s{=}(\mathcal{N},\mathcal{L})$ has nodes with CPU/memory capacity, idle power P_n^{idle}, dynamic coefficient P_n^{dyn}, hardware energy efficiency ϕ_n (J/FLOP), and grid carbon intensity $\rho_n(t)$. Links have bandwidth b_l, delay τ_l, and power coefficient μ_l^{link}. AI inference VNFs $v_i{\in}\mathcal{V}^{\text{AI}}$ specify per-request FLOPs Φ_{v_i}, enabling precise inference energy accounting. Binary placement variables $x_{v_i,n} \in \{0,1\}$ indicate whether v_i is assigned to node n; the full placement is $\mathbf{X} = [x_{v_i,n}]_{m\times|\mathcal{N}|}$, and routing is shortest-path given $\mathbf{X}$. Node energy decomposes into idle, dynamic, and AI inference terms:

$$E_n(\mathbf{X}) = P_n^{\text{idle}}\mathbb{I}[n \text{ active}] + P_n^{\text{dyn}} u_n^{\text{non-AI}}(\mathbf{X}) + \phi_n \sum_{v_i\in\mathcal{V}^{\text{AI}}} x_{v_i,n}\Phi_{v_i}\lambda_r, \tag{1}$$

where $u_n^{\text{non-AI}}$ is the CPU utilization induced by non-AI VNFs only (the AI-induced load is accounted for by the third term to avoid double counting) and λ_r is the request rate. The AI inference term steers placement toward hardware-efficient accelerators (low ϕ_n). Link energy is $E_l(\mathbf{X}) = \mu_l^{\text{link}} \sum_{(v_i,v_j)} f_{ij} p_{ij}^l$.

Multi-objective green formulation. Minimize $\mathbf{O}(\mathbf{X}) = [\mathcal{O}_1,\ldots,\mathcal{O}_5]^\top$:

$$\mathcal{O}_1 = \sum_n \phi_n \sum_{v_i\in\mathcal{V}^{\text{AI}}} x_{v_i,n}\Phi_{v_i}\lambda_r, \qquad \mathcal{O}_2 = \sum_n \big(P_n^{\text{idle}}\mathbb{I}[n \text{ active}] + P_n^{\text{dyn}} u_n^{\text{non-AI}}\big) + \sum_l E_l, \tag{2}$$

$$\mathcal{O}_3 = \sum_n \rho_n(t) E_n + \sum_l \bar{\rho}(t) E_l, \quad \mathcal{O}_4 = \sum_{v_i} \frac{d_{v_i}^{\text{CPU}}}{\alpha c_{n(v_i)}^{\text{CPU}}} + \sum_{(v_i,v_j)} \sum_{l\in\text{path}} \tau_l, \tag{3}$$

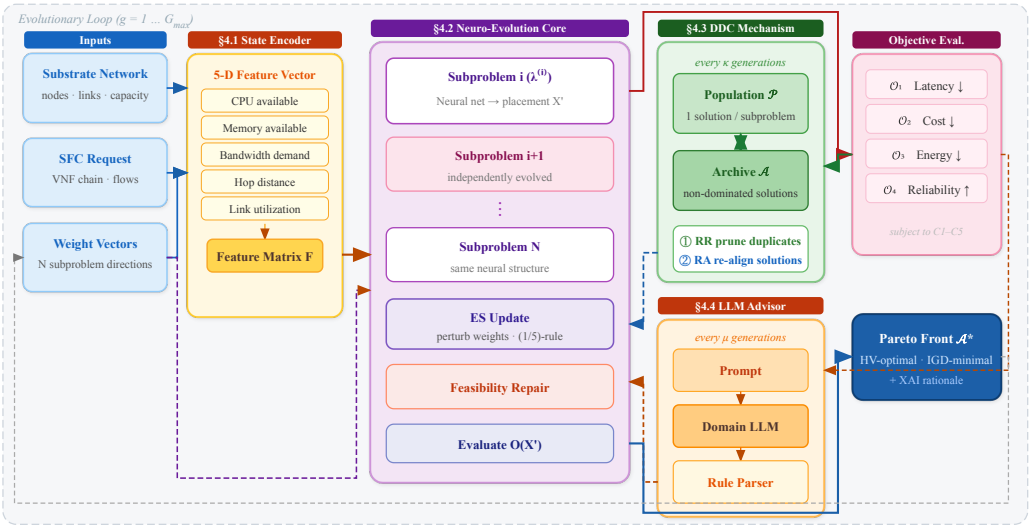


Figure 1: Overall GENO architecture. (a) State Encoder maps topology, power profiles, carbon signals, and AI FLOPs into a 7-D feature matrix $\mathbf{F}$. (b) Neuro-Evolution Core evolves N per-subproblem networks via ES, steered by $\tilde{\phi}_n$ and $\tilde{\rho}_n(t)$. (c) DDC couples population $\mathcal{P}$ with green archive $\mathcal{A}$; RR retains the lowest-$(\mathcal{O}_1{+}\mathcal{O}_3)$ representative; RA re-assigns with green-priority tie-breaking. (d) Policy Advisor (Qwen2.5-7B-Instruct-Q4, local) consumes an energy KPI summary every μ generations and emits low-carbon directives parsed into $\Delta\boldsymbol{\lambda}$ subject to a green floor $\lambda_1{+}\lambda_2{+}\lambda_3{\geq}0.5$.

$$\mathcal{O}_5 = -\log\Big(\prod_{v_i} r^{\text{node}}_{n(v_i)} \prod_{(v_i,v_j)} \prod_{l} r^{\text{link}}_{l}\Big). \tag{4}$$

$\mathcal{O}_1$ and $\mathcal{O}_2$ are additively non-overlapping: by construction $\mathcal{O}_1 + \mathcal{O}_2 = \sum_n E_n + \sum_l E_l$, i.e. total system energy decomposed into AI- and non-AI components rather than counted twice. GENO converts multi-objective to single-objective per subproblem i using the augmented Tchebycheff function

$$g_{\text{te}}(\mathbf{O}|\boldsymbol{\lambda}^{(i)}) = \max_{k=1}^{5} \lambda_k^{(i)} \big|O_k - z_k^*\big| + \delta \sum_{k=1}^{5} \lambda_k^{(i)} \big|O_k - z_k^*\big|, \tag{5}$$

with ideal point $\mathbf{z}^*$ and augmentation coefficient $\delta{=}10^{-5}$, which guarantees properly (not merely weakly) Pareto-optimal minimizers. A feasible $\mathbf{X}$ satisfies (C1) unique VNF placement; (C2–C3) CPU/memory capacity; (C4) bandwidth capacity; (C5) peak power budget $E_n(\mathbf{X})/\Delta t \leq P_n^{\max}$; (C6) SFC DAG order. The problem is isomorphic to the Many-Objective Knapsack Problem (Ishibuchi et al., 2008) and hence NP-hard (Appendix B).

4 The GENO Framework

GENO integrates four synergistic components: a State Encoder, a Neuro-Evolution Core, DDC, and a Policy Advisor (Figure 1, Algorithm 1). The state encoder feeds energy-context features to the neural operators; DDC uses these features to maintain green-biased archive diversity; and the policy advisor dynamically adjusts the weight vectors governing all subproblems.

State representation. For each candidate assignment (v_i, n), the 7-D feature vector is

$$\mathbf{f}_{v_i,n} = \big[\tilde{c}_n^{\text{CPU}},\ \tilde{c}_n^{\text{MEM}},\ \tilde{f}_i^{\text{cum}},\ \tilde{d}(n_{\text{prev}}, n),\ \tilde{\rho}_{\text{path}},\ \tilde{\phi}_n,\ \tilde{\rho}_n(t)\big]^\top, \tag{6}$$

where the two new green features are $\tilde{\phi}_n$ (normalized hardware energy-per-FLOP) and $\tilde{\rho}_n(t)$ (normalized real-time grid carbon intensity, updated each generation to track diurnal renewable availability). Stacking all $m \cdot |\mathcal{N}|$ vectors yields $\mathbf{F} \in \mathbb{R}^{m|\mathcal{N}|\times 7}$. The 7-D encoding

Algorithm 1 GENO: Energy-Efficient Network Orchestration

Input: Substrate $\mathcal{G}_s$, SFC $\mathcal{R}$, power profiles, carbon signals $\{\rho_n(t)\}$, weights $\{\boldsymbol{\lambda}^{(i)}\}$, periods κ, μ, $G_{\max}$
Output: Green Pareto front $\mathcal{A}^*$
1: Encode $\mathbf{F} \in \mathbb{R}^{m|\mathcal{N}|\times 7}$; Initialize $\mathcal{P}$ (green-first heuristic), $\mathcal{A}\leftarrow\emptyset$, $\{\boldsymbol{\theta}^{(i)}\}$
2: for $g = 1, \ldots, G_{\max}$ do
3: Update carbon signals $\rho_n(t)$
4: if $g \bmod \mu = 0$ then Query advisor $\rightarrow$ directive $\rightarrow \Delta\boldsymbol{\lambda}$; update $\boldsymbol{\lambda}^{(i)}$
5: for each subproblem i do
6: Generate offspring $\mathbf{X}' \leftarrow \mathcal{F}_{\boldsymbol{\theta}^{(i)}}(\mathbf{F}, \mathbf{X}^{(i)})$; repair to feasible; evaluate $\mathbf{O}(\mathbf{X}')$
7: Update $\mathcal{P}$ (Tchebycheff); update $\mathcal{A}$ (non-dominated); Evolve $\boldsymbol{\theta}^{(i)} \leftarrow \boldsymbol{\theta}^{(i)} + \sigma_\theta \boldsymbol{\varepsilon}$
8: if $g \bmod \kappa = 0$ then Apply DDC (RR + RA)
9: return $\mathcal{A}$ as $\mathcal{A}^*$

achieves +6.3% HV over a 5-D energy-agnostic encoding at matched (indeed fewer) parameters (Appendix B).

Neuro-evolutionary operator. Each subproblem i uses a single-hidden-layer network $\mathcal{F}_{\boldsymbol{\theta}^{(i)}}(\mathbf{f}) = \sigma(\mathbf{W}_2 \mathrm{ReLU}(\mathbf{W}_1\mathbf{f} + \mathbf{b}_1) + b_2)$ with $\mathbf{W}_1 \in \mathbb{R}^{h\times 7}$, $\mathbf{W}_2 \in \mathbb{R}^{1\times h}$, output thresholded at 0.5, and parameter count $D_\theta = 9h + 1$ (h=64). The ES update $\boldsymbol{\theta}^{(i)} \leftarrow \boldsymbol{\theta}^{(i)} + \sigma_\theta\boldsymbol{\varepsilon}$ uses (1/5)-success-rule adaptation, where a "success" is $g_{\mathrm{te}}(\mathbf{O}(\mathbf{X}') \mid \boldsymbol{\lambda}^{(i)}) < g_{\mathrm{te}}(\mathbf{O}(\mathbf{X}^{(i)}) \mid \boldsymbol{\lambda}^{(i)})$. Initialization uses a green-first heuristic (VNFs greedily assigned to the node minimizing $\mathcal{O}_1{+}\mathcal{O}_2$ subject to (C1)–(C5)), warm-starting in the energy-efficient region. A closed-loop feasibility repair enforces (C1)–(C5) with guaranteed termination (Appendix B). Forward evaluation costs $O(m|\mathcal{N}|h)$; overall complexity is $O(G_{\max}Nm|\mathcal{N}|h)$, linear in problem dimensions.

4.1 Decomposition–Dominance Coordination

DDC is applied every κ generations to prevent diversity collapse in the energy–carbon–latency subspace via two steps. (i) Redundancy Removal (RR): near-duplicate archive solutions ($\ell_\infty{<}\epsilon_{\mathrm{RR}}$) are pruned, retaining the one with lowest $\mathcal{O}_1{+}\mathcal{O}_3$—a green-biased rule preserving energy-efficient non-dominated solutions. (ii) Energy-Priority Re-Association (RA): each subproblem i is re-assigned $\arg\min_{\mathbf{X}\in\mathcal{P}\cup\mathcal{A}} g_{\mathrm{te}}(\mathbf{O}(\mathbf{X})|\boldsymbol{\lambda}^{(i)})$, with ties broken toward lower $\mathcal{O}_1$ for energy/carbon-dominant subproblems ($\sum_{k=1}^{3}\lambda_k{>}0.5$). Table 1 shows RA contributes more than RR (-9.1% vs. -5.3% HV when each removed), and their synergy is super-additive (-21.1% jointly vs. -14.4% additive). Unlike MOEA/D-AGR (Qi et al., 2014), which relocates weight vectors within the population, RA draws from $\mathcal{P}\cup\mathcal{A}$ and re-assigns existing solutions without modifying the weight structure, while RR anchors objective-space diversity independently of population behavior (Appendix B).

To characterize the failure mode, let the subproblem alignment error for subproblem i at generation g be

$$\varepsilon_i^{(g)} = g_{\mathrm{te}}\Big(\mathbf{O}(\mathbf{X}_g^{(i)}) \,\Big|\, \boldsymbol{\lambda}^{(i)}\Big) - \min_{\mathbf{X}\in\mathcal{P}_g\cup\mathcal{A}_g} g_{\mathrm{te}}\Big(\mathbf{O}(\mathbf{X}) \,\Big|\, \boldsymbol{\lambda}^{(i)}\Big), \tag{7}$$

i.e. how far a subproblem's current solution is, in scalarized value, from the best solution available for its weight vector. By definition $\varepsilon_i^{(g)}{\geq}0$, and under MOEA/D's local update it can grow unboundedly. RA drives $\varepsilon_i^{(g)}{=}0$ for all i at each invocation epoch, resetting the mis-binding that AGR's weight relocation does not in general eliminate.

4.2 Formal Guarantees

The two propositions below make the notion of search bias precise by characterizing (i) the invariance preserved by RR and (ii) the bounded influence of the preference channel. Full proofs are in Appendix C.

Table 1: DDC ablation on AI-EO (30 runs, $p<0.01$).

Variant	HV	Energy (kWh)	ΔHV
Full DDC (RR+RA)	$\mathbf{0.883 \pm 0.011}$	$\mathbf{2.67 \pm 0.13}$	—
RR only (no RA)	0.803 ± 0.019	3.24 ± 0.21	-9.1%
RA only (no RR)	0.836 ± 0.016	2.98 ± 0.18	-5.3%
No DDC	0.697 ± 0.025	4.11 ± 0.28	-21.1%

Proposition 1 (Green-region invariance of RR). Let $\mathcal{A}_g$ be the archive at generation g, $\mathcal{A}_g^\star$ the post-RR archive, and $\bar{E}_g = \frac{1}{|\mathcal{A}_g|}\sum_{\mathbf{X}\in\mathcal{A}_g}(\mathcal{O}_1(\mathbf{X}) + \mathcal{O}_3(\mathbf{X}))$ the green-weighted archive score. Then, with each retained representative weighted by the number of pre-RR solutions in its redundancy cell, RR does not increase the score: $\bar{E}_g^\star \le \bar{E}_g$, strictly whenever RR removes at least one strictly dominated near-duplicate.

Proposition 2 (Lipschitz bound on preference-channel perturbations). Let two consecutive preference epochs produce $\boldsymbol{\lambda}_g^{(i)}$ and $\boldsymbol{\lambda}_{g+\mu}^{(i)}$ subject to $|\Delta\lambda_k| \le \eta$ and the green floor $\sum_{k=1}^{3}\lambda_k \ge 1/2$. Then for every $\mathbf{X}$ and subproblem i,

$$\left|g_{\mathrm{te}}(\mathbf{O}(\mathbf{X}) \mid \boldsymbol{\lambda}_{g+\mu}^{(i)}) - g_{\mathrm{te}}(\mathbf{O}(\mathbf{X}) \mid \boldsymbol{\lambda}_g^{(i)})\right| \;\le\; 5\eta \cdot \max_k \left|\mathcal{O}_k(\mathbf{X}) - z_k^\star\right|, \tag{8}$$

and the green floor is preserved, i.e. the perturbed weight vector remains in the green half-simplex.

Proposition 1 formalizes the algorithmic side of GENO's search bias—at every DDC epoch the archive's centre of mass shifts (weakly) further into the low-energy/low-carbon region—while Proposition 2 bounds the interface side, capping any single preference perturbation at 5η times the objective-space radius.

Policy-guided weight perturbation. We expose a generic bounded weight-perturbation interface; a local LLM (Qwen2.5-7B-Instruct, 4-bit, run locally without fine-tuning) is one concrete instantiation, queried every μ=20 generations with an energy-KPI prompt and parsed into signed perturbations $\Delta\lambda_k = \eta \cdot s_k \cdot \mathbb{I}[\text{keyword}_k \in \text{directive}]$ under $|\Delta\lambda_k| \le \eta$=0.05 and the green floor $\lambda_1+\lambda_2+\lambda_3 \ge 0.5$. Any alternative perturbation source respecting Proposition 2 can be substituted; the channel accounts for only -2.8% HV in ablation, so it is a policy instantiation rather than a controller. Prompt templates, the keyword vocabulary, hallucination safeguards, and a full overhead accounting (Table 6) are in Appendix D.

5 Experiments

We evaluate GENO against seven baselines across three benchmarks to answer: RQ1 Does GENO reduce energy and carbon while preserving Pareto quality? RQ2 What is each component's contribution? RQ3–RQ4 How does GENO scale, and does optional policy guidance yield actionable, bounded improvements (Appendix E)?

Setup. (1) VNF-CD: NSFNET (14 nodes, 21 links); 3–5 VNF SFCs with 2 AI VNFs ($\Phi_{v_i} \in [10^{11}, 10^{12}]$ FLOPs). (2) AI-EO: EdgeCloudSim (Sonmez et al., 2018), 10 edge + 2 cloud nodes; 80% AI inference tasks. (3) DC-TE: Fat-tree (Al-Fares et al., 2008) (k=4, 20 nodes). Baselines: NSGA-II (Deb et al., 2002), NSGA-III (Deb & Jain, 2014), MOEA/D (Zhang & Li, 2007), MOEA/D-AGR (Qi et al., 2014), MARO (Kato et al., 2017), DeepRM (Mao et al., 2016), and GreenVNF (Liang et al., 2024) (single-objective energy minimization). All evolutionary baselines are matched at N=100, $G_{\max}$=500 (50K FEs); DRL baselines receive 50K interaction steps (verified converged). Hyperparameters: h=64, σ_θ=0.01, κ=10, $\epsilon_{\mathrm{RR}}=10^{-3}$, η=0.05. Metrics: HV (normalized), EPR (Wh/request), carbon (gCO_2/SFC), over 30 runs; pairwise Wilcoxon rank-sum tests with Holm–Bonferroni correction at α=0.01, effect sizes as Cliff's δ (†: $|\delta| \ge 0.47$). Each percentage is stated relative to the named baseline; value triples are in VNF-CD/AI-EO/DC-TE order.

Comparative results (RQ1). GENO achieves the highest HV and lowest EPR across all three benchmarks (Table 2). EPR reductions vs. MOEA/D-AGR are 37.2%/38.1%/26.4%; vs.

Table 2: Performance across three benchmarks (mean±std, 30 runs). Bold: best; underline: 2nd best. †: GENO significantly better than all baselines ($p<0.01$). EPR = energy per request (Wh).

Algorithm	VNF-CD		AI-EO		DC-TE	
	HV ↑	EPR (Wh) ↓	HV ↑	EPR (Wh) ↓	HV ↑	Energy (kWh) ↓
GENO (Ours)†	0.891±.010	1.84±.11	0.883±.011	0.52±.04	0.905±.008	2.31±.14
GreenVNF	0.712±.023	2.41±.17	0.643±.027	0.71±.06	0.758±.021	2.93±.19
MOEA/D-AGR	0.698±.025	2.93±.22	0.759±.022	0.84±.08	0.741±.023	3.14±.22
MOEA/D	0.671±.028	3.11±.24	0.736±.024	0.91±.09	0.718±.025	3.28±.24
NSGA-III	0.613±.034	3.54±.28	0.681±.031	1.08±.11	0.659±.032	3.79±.28
NSGA-II	0.581±.037	3.73±.31	0.652±.034	1.17±.12	0.631±.035	3.98±.31
MARO	0.629±.031	3.32±.26	0.704±.028	0.98±.10	0.678±.029	3.56±.27
DeepRM	0.617±.033	3.46±.29	0.693±.030	1.03±.10	0.665±.031	3.68±.29

Table 3: Individual KPI comparison on AI-EO (knee-point solution).

Algorithm	Inf. Energy (Wh,↓)	Carbon (gCO_2,↓)	Latency (ms,↓)	Reliab. (%,↑)
GENO	0.52	6.1	21.3	97.8
GreenVNF	0.71	8.4	20.1	95.9
MOEA/D-AGR	0.84	10.2	22.6	96.5
MOEA/D	0.91	10.8	23.4	96.1
MARO	0.98	11.9	26.3	95.3
DeepRM	1.03	12.4	27.1	95.1

GreenVNF, 23.7%/26.8%/21.2%. HV gains over MOEA/D-AGR are 27.6%/16.3%/22.1% ($p<0.01$). Aggregating the AI-EO knee-point KPIs (Table 3), none of the seven baselines simultaneously enters the corner defined by EPR ≤ 0.6 Wh/request and tail latency ≤ 22 ms: GreenVNF passes latency (20.1 ms) but fails energy (0.71 Wh); MOEA/D-AGR fails both (0.84 Wh, 22.6 ms). GENO reaches this corner (0.52 Wh, 21.3 ms)—a 26.8% EPR improvement over GreenVNF and 38.1% over MOEA/D-AGR at comparable latency—while also achieving the lowest carbon and highest reliability. We report this corner as one representative low-energy operating point rather than a claim of general superiority. GENO also attains the lowest IGD on all three benchmarks (0.041/0.038/0.035 vs. 0.112/0.094/0.108 for MOEA/D-AGR), and reaches 90% of its final HV by 16K FEs (Appendix E).

Component ablation (RQ2). Table 4 decomposes GENO. DDC is most critical (−21.1% HV, +157.7% EPR): without diversity maintenance the optimizer collapses to high-energy local optima. The AI energy model (+75% EPR when removed) and the green encoder features $\tilde{\phi}_n, \tilde{\rho}_n$ (+86.5% EPR) are the next most important: ignoring inference FLOPs or hardware/carbon signals places VNFs on CPUs rather than efficient accelerators. The carbon objective cannot be proxied by energy minimization (+30.8% carbon without $\mathcal{O}_3$). Policy guidance provides a bounded gain: "Trad. Adaptive Weighting" (−3.9% HV / +25.0% EPR) and "w/o LLM Advisor" (−2.8% HV / +21.2% EPR) confirm the two are genuinely distinct interventions, with dominant gains attributable to the core optimization components.

6 Analysis and Discussion

Why DDC provides the largest gains. The low-energy Pareto frontier lies far from both the initial feasible population and the latency-efficient region where standard MOEAs converge. DDC's RA step bridges this gap by re-assigning each subproblem to the best solution in $\mathcal{P}\cup\mathcal{A}$, exploiting the dominance archive beyond neighborhood-restricted updates, while RR keeps that archive uncluttered; their super-additive interaction (−21.1% vs. −14.4% additive) confirms complementarity. Energy vs. carbon. Removing $\mathcal{O}_3$ raises realized carbon by 30.8% with only 2.5% HV reduction, showing energy and carbon are not interchangeable: energy minimization alone may select low-PUE but coal-heavy grids. Robustness. The DDC-driven HV and diversity gains are structural properties of the search and in-

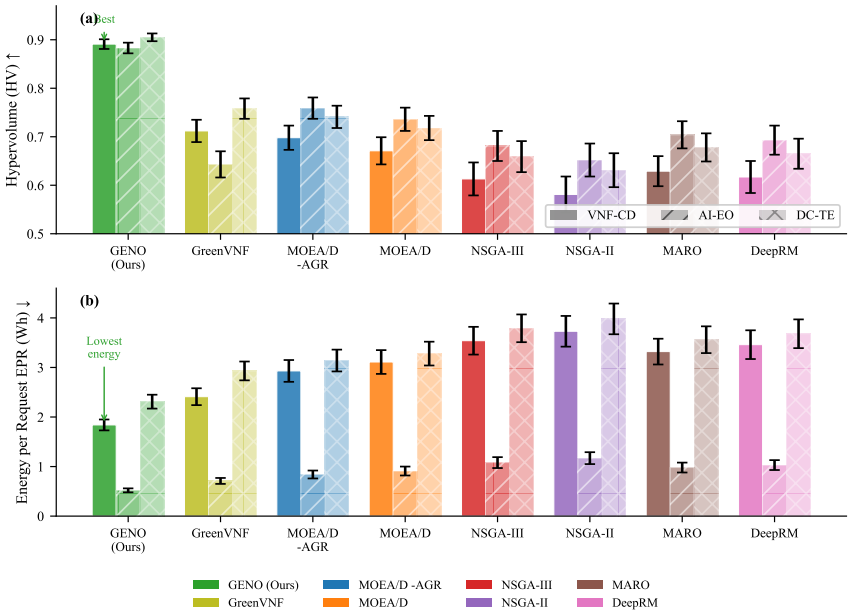


Figure 2: HV (↑, top) and EPR (Wh, ↓, bottom) across three benchmarks (30 runs, ±1 std). GENO achieves the strongest combined HV/EPR profile among the compared methods.

Table 4: Component ablation on AI-EO (30 runs, all $p<0.01$). Relative changes vs. full GENO.

Variant	HV	EPR (Wh)	ΔHV	ΔEPR
GENO (Full)	0.883	0.52	—	—
Trad. Adaptive Weight.	0.849	0.65	−3.9%	+25.0%
w/o $\mathcal{O}_3$ carbon	0.861	0.68	−2.5%	+30.8%
w/o AI energy model	0.834	0.91	−5.5%	+75.0%
w/o Energy features	0.801	0.97	−9.3%	+86.5%
w/o DDC	0.697	1.34	−21.1%	+157.7%
w/o LLM Advisor	0.858	0.63	−2.8%	+21.2%

dependent of measurement noise, whereas the carbon reductions are the most exposed to forecast error—though a carbon-noise sensitivity analysis shows they degrade gracefully up to σ_ρ=25% (Appendix E).

Limitations. (i) All experiments are simulation-based; the carbon traces and workloads are synthetic, so absolute energy/carbon figures should be read as relative comparisons under a shared model rather than operational forecasts. (ii) Carbon signals are historical rather than real-time grid APIs (Radovanović et al., 2023). (iii) For $|\mathcal{N}|>200$, advisor prompt length grows linearly and requires hierarchical summarization. (iv) Water-usage effectiveness is a natural sixth objective (Li et al., 2023). The framework is also applicable to O-RAN (Polese et al., 2023), edge split inference (Kang et al., 2017), and carbon-aware 6G slicing (Foukas et al., 2017).

7 Conclusion

We presented GENO, a multi-objective green optimization framework for AI service orchestration. By treating inference energy, network energy, and carbon as first-class objectives and aligning these signals across the state representation, the search mechanism, and an optional policy channel, GENO reaches a low-energy operating point unattained by the evaluated baselines within the same compute budget, while preserving competitive latency

and reliability. The results indicate that holistic multi-objective green optimization—rather than single-metric energy minimization—is a viable design principle for sustainable AI service platforms. Future work includes carbon-temporal scheduling, distilling the advisor into an edge-deployable module, and net-zero service certification with formal carbon-footprint bounds.

Reproducibility Statement

All source code (the GENO optimizer, the three benchmark generators, the synthetic carbon-trace generator, the LLM advisor prompts and parser, and the experiment-driver scripts) will be released under a permissive open-source license, together with the raw per-run logs (30 seeds $\times$ 3 benchmarks $\times$ 8 algorithms), random seeds, hyperparameter manifests, and hardware configuration needed to reproduce every table and the ablations. Full method details are in Appendix B and proofs in Appendix C.

References

Mohammad Al-Fares, Alexander Loukissas, and Amin Vahdat. A Scalable, Commodity Data Center Network Architecture. ACM SIGCOMM Computer Communication Review, 38 (4):63–74, 2008. doi: 10.1145/1402958.1402967.

Stefano Buzzi, Chih-Lin I, Thierry E. Klein, H. Vincent Poor, Chenyang Yang, and Alessio Zappone. A Survey of Energy-Efficient Techniques for 5G Networks and Challenges Ahead. IEEE Journal on Selected Areas in Communications, 34(4):697–709, 2016. doi: 10.1109/JSAC.2016.2550338.

Kalyanmoy Deb and Himanshu Jain. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. IEEE Transactions on Evolutionary Computation, 18 (4):577–601, 2014. doi: 10.1109/TEVC.2013.2281535.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation, 6(2):182–197, 2002. doi: 10.1109/4235.996017.

Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. Trends in ai inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. Sustainable Computing: Informatics and Systems, 38:100857, 2023.

Xenofon Foukas, Georgios Patounas, Ahmed Elmokashfi, and Mahesh K. Marina. Network Slicing in 5G: Survey and Challenges. IEEE Communications Magazine, 55(5):94–100, 2017. doi: 10.1109/MCOM.2017.1600951.

Songtao Guo, Junyi Liu, Yuanyuan Yang, Bin Xiao, and Zhi Li. Energy-Efficient Dynamic Computation Offloading and Cooperative Task Scheduling in Mobile Cloud Computing. IEEE Transactions on Mobile Computing, 18(2):319–333, 2018. doi: 10.1109/TMC.2018.2831230.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. arXiv preprint arXiv:2409.12186, 2024.

Hisao Ishibuchi, Noritaka Tsukamoto, and Yusuke Nojima. Evolutionary Many-Objective Optimization: A Short Review. In IEEE Congress on Evolutionary Computation (CEC 2008), pp. 2419–2426, 2008. doi: 10.1109/CEC.2008.4631121.

Yiping Kang, Johann Hauswald, Cao Gao, Austin Rovinski, Trevor Mudge, Jason Mars, and Lingjia Tang. Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge. ACM SIGPLAN Notices, 52(4):615–629, 2017. doi: 10.1145/3093336.3037698.

Nei Kato, Zubair Fadlullah, Bowen Mao, Fengxiao Tang, Osamu Akashi, Takeru Inoue, and Kimihiro Mizutani. The Deep Learning Vision for Heterogeneous Network Traffic Control: Proposal, Challenges, and Future Perspective. IEEE Wireless Communications, 24(3):146–153, 2017. doi: 10.1109/MWC.2016.1600317WC.

Atefeh Khosravi, Saurabh Kumar Garg, and Rajkumar Buyya. Energy and Carbon-Efficient Placement of Virtual Machines in Distributed Cloud Data Centers. In Euro-Par 2013 Parallel Processing, Lecture Notes in Computer Science, pp. 317–328, 2013. doi: 10.1007/978-3-642-40047-6_33.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the Carbon Emissions of Machine Learning. arXiv preprint arXiv:1910.09700, 2019.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. Green Algorithms: Quantifying the Carbon Footprint of Computation. Advanced Science, 8(12):2100707, 2021. doi: 10.1002/advs.202100707.

Pengfei Li, Jianyi Rao, Geoffrey Hasegawa, Zhen Zhao, Zheng Zheng, and Yan Liu. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. arXiv preprint arXiv:2304.03271, 2023.

Junbin Liang, Shaodong Huang, Yu Qiu, Lu Liu, Furqan Aziz, and Min Chen. Sustainable virtual network function placement and traffic routing for green mobile edge networks. IEEE Transactions on Green Communications and Networking, 8(4):1450–1465, 2024.

David López-Pérez, Antonio De Domenico, Nicola Piovesan, Xinli Geng, Harvey Bao, Qitao Song, and Mérouane Debbah. A Survey on 5G Radio Access Network Energy Efficiency: Massive MIMO, Lean Carrier Design, Sleep Modes, and Machine Learning. IEEE Communications Surveys & Tutorials, 24(1):653–697, 2022. doi: 10.1109/COMST.2022.3142532.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. Journal of Machine Learning Research, 24(253):1–15, 2023.

Bomin Mao, Fengxiao Tang, Yuichi Kawamoto, and Nei Kato. Ai models for green communications towards 6g. IEEE Communications Surveys & Tutorials, 24(1):210–247, 2021.

Hongzi Mao, Mohammad Alizadeh, Ishai Menache, and Srikanth Kandula. Resource Management with Deep Reinforcement Learning. In Proceedings of the 15th ACM Workshop on Hot Topics in Networks (HotNets '16), pp. 50–56, 2016. doi: 10.1145/3005745.3005750.

Yuyi Mao, Changsheng You, Jun Zhang, Kaibin Huang, and Khaled B. Letaief. A Survey on Mobile Edge Computing: The Communication Perspective. IEEE Communications Surveys & Tutorials, 19(4):2322–2358, 2017. doi: 10.1109/COMST.2017.2745201.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon Emissions and Large Neural Network Training. arXiv preprint arXiv:2104.10350, 2021.

Michele Polese, Leonardo Bonati, Salvatore D'Oro, Stefano Basagni, and Tommaso Melodia. Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges. IEEE Communications Surveys & Tutorials, 25(2):1376–1411, 2023. doi: 10.1109/COMST.2023.3239220.

Yutao Qi, Xiaoliang Ma, Fang Liu, Licheng Jiao, Jianyong Sun, and Jianshe Wu. MOEA/D with Adaptive Weight Adjustment. Evolutionary Computation, 22(2):231–264, 2014. doi: 10.1162/EVCO_a_00109.

Asfandyar Qureshi, Rick Weber, Hari Balakrishnan, John Guttag, and Bruce Maggs. Cutting the Electric Bill for Internet-Scale Systems. In Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication, pp. 123–134, 2009. doi: 10.1145/1592568.1592584.

Ana Radovanović, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyue Xiao, Maya Haridasan, et al. Carbon-Aware Computing for Datacenters. IEEE Transactions on Power Systems, 38(2):1270–1280, 2023. doi: 10.1109/TPWRS.2022.3173250.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. Communications of the ACM, 63(12):54–63, 2020. doi: 10.1145/3381831.

Cagatay Sonmez, Atay Ozgovde, and Cem Ersoy. EdgeCloudSim: An Environment for Performance Evaluation of Edge Computing Systems. Transactions on Emerging Telecommunications Technologies, 29(11):e3493, 2018. doi: 10.1002/ett.3493.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 3645–3650, 2019. doi: 10.18653/v1/P19-1355.

Philipp Wiesner, Ilja Behnke, Dominik Scheinert, Kordian Gontarska, and Lauritz Thamsen. Let's Wait Awhile: How Temporal Workload Shifting Can Reduce Carbon Emissions in the Cloud. In Proceedings of the 22nd International Middleware Conference, pp. 260–272, 2021. doi: 10.1145/3464298.3493399.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Behrooz, James Huang, Yue Bai, et al. Sustainable AI: Environmental Implications, Challenges, and Opportunities. In Proceedings of Machine Learning and Systems (MLSys), volume 4, pp. 795–813, 2022.

Ke Zhang, Yuming Mao, Supeng Leng, Quanxin Zhao, Longjiang Li, Xin Peng, Li Pan, Sabita Maharjan, and Yan Zhang. Energy-efficient offloading for mobile edge computing in 5g heterogeneous networks. IEEE access, 4:5896–5907, 2016.

Qingfu Zhang and Hui Li. MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition. IEEE Transactions on Evolutionary Computation, 11(6):712–731, 2007. doi: 10.1109/TEVC.2007.892759.

A Notation

Table 5: Summary of key notation.

Symbol	Description
$\mathcal{G}_s = (\mathcal{N}, \mathcal{L})$	Substrate network graph
$\mathcal{R} = (\mathcal{V}, \mathcal{E})$	SFC request (VNFs + flows)
$m = \lvert\mathcal{V}\rvert$	Number of VNFs in SFC
$c_n^{\mathrm{CPU}}, c_n^{\mathrm{MEM}}$	CPU/memory capacity of node n
b_l	Bandwidth capacity of link l
$P_n^{\mathrm{idle}}, P_n^{\mathrm{dyn}}$	Idle and dynamic power of node n
ϕ_n	Energy-per-FLOP of node n (J/FLOP)
$\rho_n(t)$	Grid carbon intensity at n ($\mathrm{gCO_2}$/kWh)
Φ_{v_i}	FLOPs/request for AI VNF v_i
μ_l^{link}	Link power per bandwidth (W/Gbps)
$x_{v_i,n} \in \{0,1\}$	Placement decision variable
$\mathbf{O} = [\mathcal{O}_1, \ldots, \mathcal{O}_5]^\top$	Five-objective green vector
$\boldsymbol{\lambda} \in \mathbb{R}^5_{\geq 0},\ \lVert\boldsymbol{\lambda}\rVert_1 = 1$	Scalarization weight vector
N	Population size

B Additional Method Details

Complexity / MaOKP isomorphism. The five-objective problem is structurally isomorphic to the Many-Objective Knapsack Problem (MaOKP) (Ishibuchi et al., 2008): each candidate assignment (v_i, n) is an item, resource demands are weights, node/link capacities are knapsack capacities, and the five objective values are negated item profits. This confirms NP-hardness and explains why greedy approaches fail: the energy–latency trade-off requires Pareto-aware search.

Objective remarks. In our benchmarks, reliability parameters are tightly concentrated (0.95–0.999), so $\mathcal{O}_5$ acts as a secondary tie-breaker among near-equivalent energy solutions while preventing reliability collapse in dominated regions. Constraint (C5) is motivated by datacenter power capping: nodes exceeding their peak power budget trigger thermal throttling, so encoding it as a hard constraint prevents thermally unsustainable placements.

Closed-loop feasibility repair. Applied after thresholding and before evaluation, the repair enforces (C1)–(C5) in two stages. First, a column-wise projection restores unique placement: for each VNF v_i, only the highest-scoring node is retained. Second, if the assignment violates CPU/memory/bandwidth/power constraints, GENO iteratively removes the assigned pair (v_i, n) with the worst energy-efficiency ratio $E_n(\mathbf{X}')/d_{v_i}^{\mathrm{CPU}}$, updates residual capacities, and reassigns v_i to the best remaining feasible node; if none exists, the offspring is rejected and the parent retained. Letting $\Psi(\mathbf{X}')$ be the number of violated constraints after enforcing unique placement, each repair step strictly reduces Ψ; since $\Psi \in \mathbb{N}_{\geq 0}$ is bounded below, the loop terminates finitely.

Energy-aware architecture ablation. On AI-EO (30 runs, $G_{\max}$=500), the 7-D encoding with h=64 attains HV 0.883 at 2.67 kWh, whereas the 5-D energy-agnostic baseline at the same width drops to HV 0.801 at 3.82 kWh (−9.3% HV). The comparison is parameter-controlled: the 7-D model with h=32 uses fewer parameters than the 5-D baseline (D_θ=321 vs. 513) yet still attains +5.7% HV and −16.5% EPR, isolating the contribution of $\tilde{\phi}_n$ and $\tilde{\rho}_n(t)$. Widening to h=128 yields only +0.2% HV.

Hyperparameters. κ=10, σ_θ=0.01 balanced over-frequent intervention against under-correction; the advisor period is μ=20=2κ so each DDC invocation uses the most recent weight vectors. The ES step floor is $\sigma_\theta^{\min}$=10^{-4}.

DDC vs. MOEA/D-AGR. Three structural differences explain DDC's advantage: (1) candidate pool—RA draws from $\mathcal{P} \cup \mathcal{A}$, exploiting non-dominated solutions overlooked by the neighborhood update; (2) repair vs. relocation—AGR moves weight vectors (changing what is optimized), while DDC re-assigns existing solutions; (3) diversity anchor—RR explicitly

maintains objective-space diversity in $\mathcal{A}$ regardless of population behavior. Energy-Priority RA additionally biases the search toward green operating points within the Pareto-equivalent set without distorting the decomposition geometry.

C Proofs

Proof of Proposition 1. RR partitions $\mathcal{A}_g$ into ℓ_∞-neighborhood cells of diameter $< \epsilon_{\text{RR}}$ and, within each cell, retains the solution with the smallest $\mathcal{O}_1+\mathcal{O}_3$. Weighting each retained representative by the occupancy of its cell, the retained value is no larger than the cell mean it replaces; summing over cells bounds the occupancy-weighted post-RR mean by the pre-RR mean $\bar{E}_g$. The inequality is strict whenever some cell holds members with distinct $\mathcal{O}_1 + \mathcal{O}_3$ values, i.e. RR removes a strictly dominated near-duplicate. The occupancy weighting is necessary: an unweighted mean can rise when RR deletes many high-score duplicates from a single cell. □

Proof of Proposition 2. Both the max term and the augmentation term of Eq. equation 5 are 1-Lipschitz in $\boldsymbol{\lambda}$ under the sup norm scaled by $\max_k |\mathcal{O}_k - z_k^\star|$. With $|\Delta\lambda_k| \le \eta$ over five objectives, $\|\boldsymbol{\lambda}_{g+\mu} - \boldsymbol{\lambda}_g\|_1 \le 5\eta$, giving the stated bound. The floor is preserved because each $\Delta\boldsymbol{\lambda}$ from the parser is projected onto the half-simplex $\{\boldsymbol{\lambda} : \sum_{k=1}^{3} \lambda_k \ge 1/2,\ \|\boldsymbol{\lambda}\|_1 = 1\}$ before it is applied. Tightness. The $5\eta R$ figure ($0.25R$ at η=0.05) is the worst case in which all five weights move by the full η; a single keyword directive moves one weight, perturbing the scalarization by at most ηR=$0.05R$. The bound is thus loose by 2.5–5× relative to the observed effect but is the correct quantity for a worst-case safety guarantee on any single directive. □

D Policy-Guided Interface Details

Prompt. Every μ=20 generations GENO queries Qwen2.5-7B-Instruct (Hui et al., 2024) (4-bit Q4_K_M, llama.cpp, local; temperature 0.2) with a green-status prompt containing current energy KPIs, per-node carbon intensity and GPU utilization, hardware efficiency ϕ_n, and SLA violation counts. The model responds "Reduce [inference energy/network energy/carbon/both] by [strategy] because [reason]." Parsing. A rule-based parser maps an energy-extended vocabulary (k=1: inference energy, GPU, FLOP efficiency; k=2: network energy, link power; k=3: carbon, emissions, renewable; k=4: latency, delay, SLA; k=5: reliability) to $\Delta\lambda_k$; initial weights $\boldsymbol{\lambda}$=$[0.30, 0.25, 0.20, 0.15, 0.10]$. Safeguards. Bounded perturbations $|\Delta\lambda_k| \le 0.05$; green floor $\lambda_1+\lambda_2+\lambda_3 \ge 0.5$; infrequent invocation (μ=20). The floor restricts the search to the green-prioritized half of the weight simplex; operators requiring the full front set it to zero.

Table 6: LLM advisor overhead per run ($G_{\max}$=500, μ=20: 25 calls). Qwen2.5-7B-Instruct-Q4 on RTX 4090. Per-call energy is board power (mean nvidia-smi reading over the call window) × latency.

Metric	Value
LLM model	Qwen2.5-7B-Instruct-Q4 (local)
Calls per run	25
Latency per call (ms)	490±110
Total LLM time (s)	12.3
Evolutionary loop time (s)	44.0
LLM time fraction	21.8%
Avg. GPU board power during call (W)	≈145
LLM energy per call (Wh)	0.020
Total LLM energy per run (Wh)	0.49
Network energy saved (Wh/req., vs. GreenVNF)	0.32
Workload-amortized benefit (per 10^4 req./window)	$\approx 6.5\times10^3\times$
Est. LLM CO_2 per run (gCO_2)	0.23

The 25 calls per run add 12.3 s and consume ≈ 0.49 Wh (≈ 0.23 gCO_2), derived from a mean board power of ≈ 145 W over each 490 ms call window—more than three orders of

magnitude below the per-window network-energy saving relative to GreenVNF (≈ 3.2 kWh over 10^4 requests).

E Additional Results

Convergence. Figure 3 shows GENO reaches 90% of its final HV by 16K FEs and surpasses GreenVNF's final EPR by 8K FEs. DDC effect. Figure 4 visualizes population evolution in the energy–carbon subspace: with DDC the population spreads toward the Pareto front, whereas without DDC it collapses by generation 100.

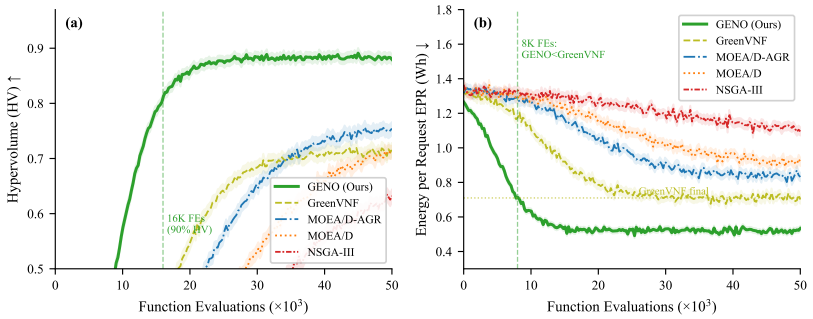


Figure 3: HV and EPR vs. function evaluations on AI-EO. GENO reaches 90% of its final HV by 16K FEs—a $> 3\times$ energy-convergence speedup vs. energy-unaware baselines. Shaded: ± 1 std, 30 runs.

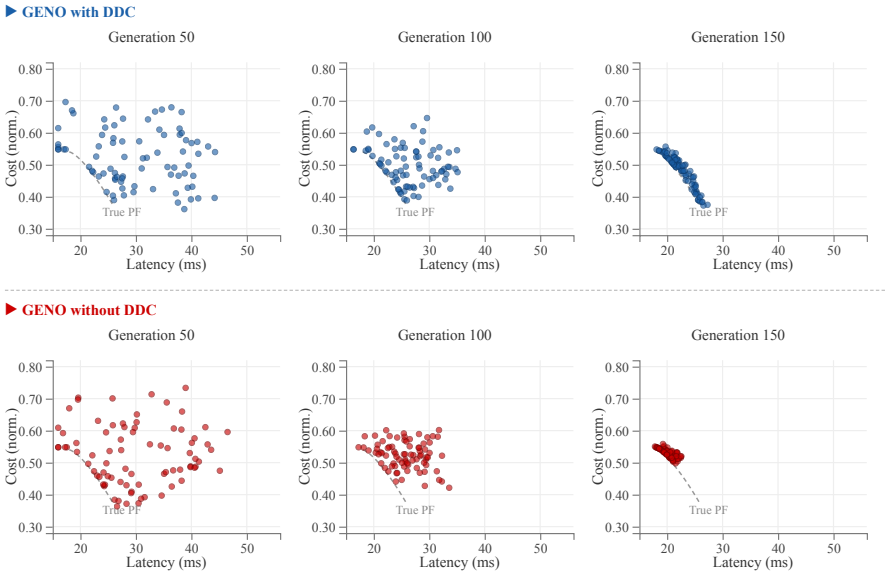


Figure 4: Population in inference-energy vs. carbon subspace at gen. 50/100/150. Top: with DDC—progressive spread toward the Pareto front. Bottom: without DDC—collapse by gen. 100.

Scalability (RQ3). HV degrades gracefully ($< 5\%$, 10→100 nodes) and EPR increases modestly (+37%); a $10\times$ AI-workload increase raises EPR by $3.5\times$ while HV drops only 0.5% (Table 7, Figure 5). At 100 nodes, 429 s remains within re-optimization windows; warm-start recovers a competitive front in ≈ 22 s.

Table 7: Scalability on AI-EO: HV and EPR vs. network size ($\Phi_{\max}$=10^{11} FLOPs unless noted).

$\lvert\mathcal{N}\rvert$	m	$\Phi_{\max}$	HV	EPR (Wh)	Time (s)
10	4	10^{11}	0.883	0.52	44
10	4	10^{12}	0.879	1.83	46
30	8	10^{11}	0.871	0.57	104
50	10	10^{11}	0.858	0.63	196
100	15	10^{11}	0.841	0.71	429

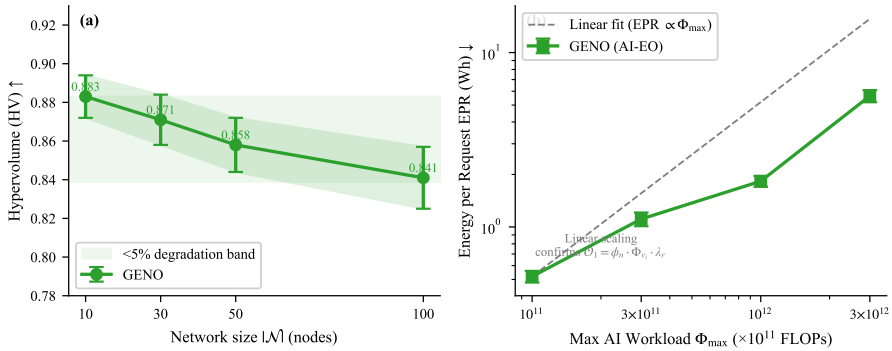


Figure 5: Left: HV vs. $|\mathcal{N}|$ (graceful $<$ 5% degradation). Right: EPR vs. AI workload $\Phi_{\max}$, showing the expected increase in energy with heavier inference load.

Policy-guidance case study (RQ4). In an adverse AI-EO instance (EPR = 1.24 Wh/request, ρ_{avg}=380 gCO_2/kWh, top-decile in the pool), the advisor proposed migrating inference VNFs from a coal-grid CPU node (ρ=510) to a wind-grid GPU node (ρ=18, ϕ=0.31 J/TFLOP) at 29% utilization. Applying $\Delta\lambda_1$=$\Delta\lambda_3$=+0.05 to energy-dominant subproblems, per-request inference energy dropped 1.24→0.61 Wh (−50.8%) and carbon 13.4→2.1 gCO_2 (−84.3%) over 20 generations, with latency rising only 2.1 ms. This instance is not typical: because most instances already sit near the green region, the channel rarely binds, consistent with its small aggregate effect (−2.8% HV / +21.2% EPR). Over the ten highest-carbon instances it reduces per-request carbon by a median of 31% (IQR 22–58%), with −84.3% the upper extreme.

Carbon sensitivity. Injecting zero-mean Gaussian noise $\sigma_\rho \in \{5\%, 15\%, 25\%\}$ into $\tilde{\rho}_n(t)$ and the $\mathcal{O}_3$ evaluation: at σ_ρ=15% realized carbon rises 8.2% and HV drops 1.4%; at σ_ρ=25% carbon degrades 14.1% yet GENO still beats GreenVNF by 21.4% on EPR. Testbed outlook. Preliminary micro-benchmarks with Llama-2-7B on a Jetson Orin AGX (65 W TDP) yield ϕ_n≈0.41 J/TFLOP, within the VNF-CD parameter range, suggesting simulation results translate to real deployments.