Quality-Diversity LLM for Generative Design

Ariq Koh [*1] **Melvin Wong** [*1] **Jiao Liu** [1] **Caishun Chen** [2] **Yew Soon Ong** [1 2]

[*]Equal contribution [1]*Nanyang Technological University, Singapore* [2]*Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore*. Correspondence to: Melvin Wong wong1357@ntu.edu.sg.

1. Introduction

Early-stage engineering design requires practitioners to navigate a vast space of candidate geometries, each subject to various functional requirements such as structural integrity and spatial constraints. Before converging on a final design specification, engineers must survey this space broadly enough to understand the tradeoffs among feasible alternatives.

Recently, text-to-X generative models have enabled the synthesis of diverse design modalities from natural-language prompts, opening opportunities for systematic exploration of engineering design spaces [1]. However, translating desirable physical properties into effective prompts is non-trivial, as the relationship between prompt phrasing and resulting geometry is complex and unknown. Quality-Diversity (QD) optimization [2] offers a principled framework for structured exploration by distributing search pressure across an entire target feature space rather than collapsing to a single optimum.

To this end, we present QD-LLMs, a novel framework that allows practitioners to rapidly survey a broad range of physical design alternatives without manually crafting prompts. The framework leverages LLMs within three specialized *natural-language-coded* emitters with complementary parent selection and variation strategies, dynamically coordinated by an LLM allocator to maximize search performance.

2. Method

Problem Formulation. Using a genetics inspired representation, for a given target domain τ, a genotype $z \in \mathcal{Z}$ is encoded as a natural language prompt, and the corresponding phenotype $x \sim G(\cdot|z)$ is a 3D geometric design synthesized by a text to 3D generative model G. Fitness combines domain alignment and physical performance: $f(x) = \lambda \cdot f_{\text{domain}}(x|\tau) + (1-\lambda) \cdot f_{\text{physical}}(x)$, where f_{domain} is evaluated by a vision-language model and f_{physical} by a domain-specific evaluator. Each phenotype is also characterized by a feature descriptor $m(x) \in \mathbb{R}^D$ that captures its physical properties (e.g., volume, center of gravity). A D-dimensional target feature space is uniformly partitioned into a grid archive $\mathcal{A}$, where each occupied cell stores a genotype-phenotype pair (z, x) (or elite). The goal is to maximize aggregate fitness across occupied cells.

Framework. Fig. 1 illustrates the proposed QD-LLMs framework. At each iteration t, an LLM-based allocator selects N LLM emitters from a pool $\mathcal{E}$ of three specialized LLM emitters based on archive coverage history and per-emitter-type success rates over a sliding window, enabling phase-dependent adaptation throughout search. Each selected emitter then generates offspring genotypes via one of the three following complementary strategies. The *Uniform LLM Emitter* (QD-LLM/UE) selects parent elites uniformly at random from occupied cells for broad exploration and generates a random variation using an LLM. The *Sparse LLM Emitter* (QD-LLM/SE) biases parent selection toward cells with few occupied neighbors, prioritizing near-neighbor expansion and generates a random variation using an LLM. Lastly, the *Directional LLM Emitter* (QD-LLM/DE) identifies the most underexplored region, selects a source elite, retrieves a guide elite along the source-to-target direction, and instructs the LLM to extrapolate beyond the guide toward the target. The resulting offspring genotypes are synthesized into corresponding phenotypes via a text-to-3D model, evaluated for domain alignment and physical performance, and inserted into the archive following an elite preservation rule.

3. Experiments and Results

We evaluate QD-LLMs on a vehicle design scenario searching for diverse, high-drag-performing cars across a 2D feature space (volume × vertical center of gravity) with a 20×20 grid archive. GPT-4o-mini serves as the LLM, Shap-E [3] as the generative model, and λ=0.3. We employ a drag surrogate model, trained following [4], to evaluate physical performance. Each method runs for 2K iterations with N=5 offspring per iteration, averaged over three runs. Our metrics are based on the Domain and Physical Alignment Rating (DPAR) introduced in [5] which penalizes methods that achieve low drag through domain invalid designs. We compare DPAR, QD-Score (aggregate DPAR across occupied archive cells), and domain score, averaged across all iterations. Table 1 shows that QD-LLMs achieves 7.4× higher DPAR (0.920 vs. 0.124), 13.8× higher domain score (0.977 vs. 0.071) and 9.2× higher QD-Score (0.883 vs. 0.096) than CMA-ME, the QD baseline described in [6]. Interestingly, as shown in Fig. 2, state-of-the-art CMA-ME commonly used in real-coded QD fills the archive but with domain-invalid designs that achieve low drag through trivial solutions (such as flat discs) rather than meaningful domain-aligned physical geometries, in contrast with our proposed QD-LLMs.

Table 1: Performance comparison (averaged).

Metric	QD-LLMs	CMA-ME
DPAR	**0.920 ± 0.025**	0.124 ± 0.053
QD-Score	**0.883 ± 0.021**	0.096 ± 0.046
Domain Score	**0.977 ± 0.028**	0.071 ± 0.032

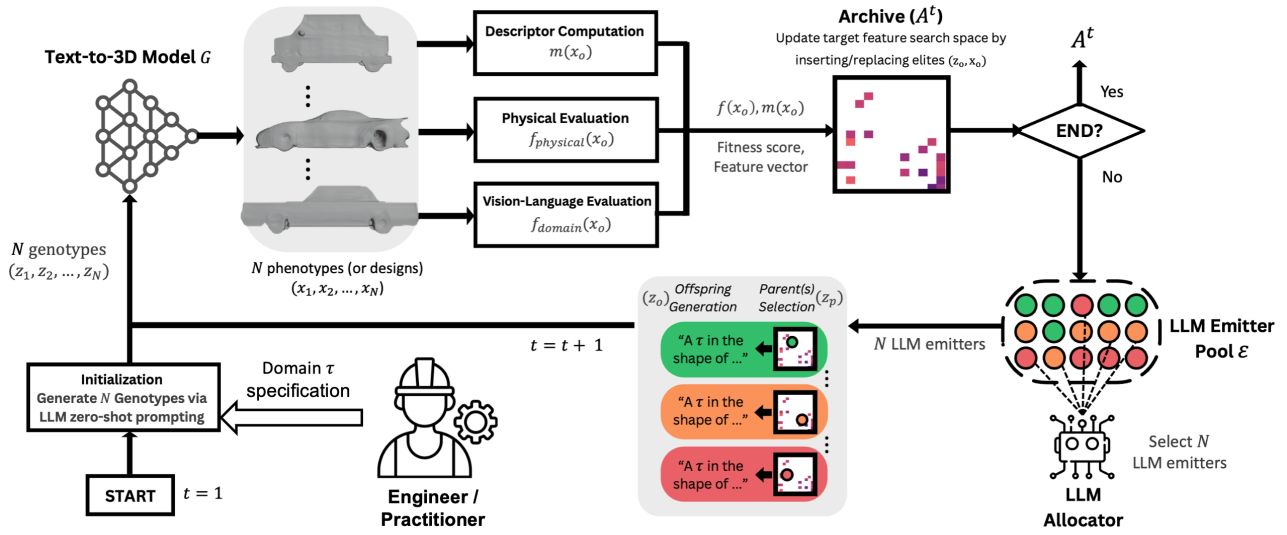

Fig. 1: Overview of the QD-LLMs framework. An engineer specifies a target domain τ, and the system initializes N genotypes. At each iteration t, a text-to-3D model G synthesizes phenotypes (3D geometries $x_1, x_2, \ldots, x_N$), which are evaluated for physical performance $f_{\text{physical}}(x_o)$, domain alignment $f_{\text{domain}}(x_o)$, and feature descriptors $m(x_o)$. Fitness scores and feature vectors are used to update the grid archive $\mathcal{A}^t$ by inserting or replacing elites (z_o, x_o). An LLM Allocator selects N emitters from the pool $\mathcal{E}$, each of which performs parent selection (z_p) and offspring generation (z_o). The loop repeats until a termination condition is met.

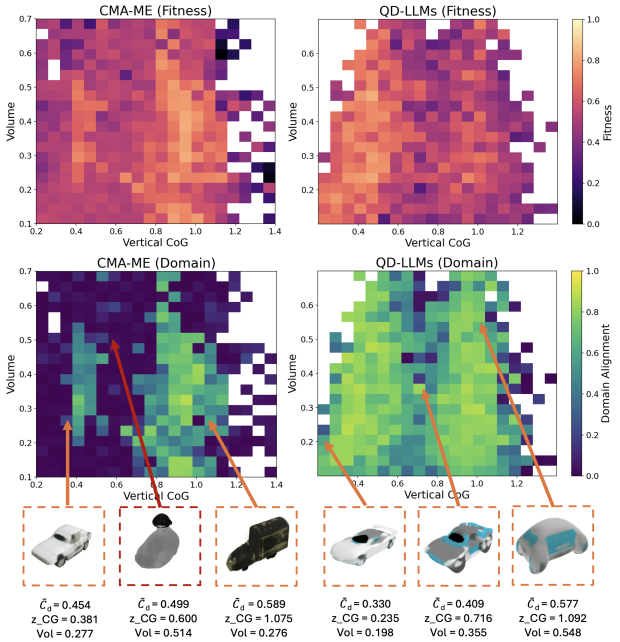

Fig. 2: Archive heatmaps comparing CMA-ME and QD-LLMs. Valid designs outlined in orange, invalid in red.

4. Conclusion

We describe QD-LLMs, a QD framework with *natural-language-coded* LLM Emitters for structured exploration of text-to-X generative design spaces. Experiments on vehicle design demonstrate over 7× improvement in domain-valid physical performance compared to classical QD methods. For engineers and scientists, QD-LLMs represents a shift from manual, intuition-driven prompt engineering to systematic, AI-driven exploration, enabling practitioners to rapidly discover high-performing design alternatives, uncover tradeoffs across physical properties, and make more informed decisions earlier in the engineering process. As text-to-X generative models continue to proliferate across engineering and scientific domains, QD-LLMs provides a ready framework for autonomous engineering design exploration.

References

[1] Melvin Wong, Jiao Liu, Thiago Rios, Stefan Menzel, and Yew-Soon Ong. Llm2tea: An agentic ai designer for discovery with generative evolutionary multitasking. *IEEE Computational Intelligence Magazine*, 20(4):42–55, 2025.

[2] Antoine Cully and Yiannis Demiris. Quality and diversity optimization: A unifying modular framework. *IEEE Transactions on Evolutionary Computation*, 22(2):245–259, 2018.

[3] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

[4] Binyang Song, Chenyang Yuan, Frank Permenter, Nikos Arechiga, and Faez Ahmed. Data-driven car drag prediction with depth and normal renderings. *Journal of Mechanical Design*, 146(5):051714, 2024.

[5] Melvin Wong, Yilin Lyu, Thiago Rios, Stefan Menzel, and Yew-Soon Ong. Llm-to-phy3d: Physically conform online 3d object generation with llms. *arXiv preprint arXiv:2506.11148*, 2025.

[6] Matthew C. Fontaine, Julian Togelius, Stefanos Nikolaidis, and Amy K. Hoover. Covariance matrix adaptation for the rapid illumination of behavior space. In *Proceedings of the 22nd Annual Conference on Genetic and Evolutionary Computation (GECCO)*, pages 94–102, 2020.