Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

QD-LLMs: Quality-Diversity Optimization with LLMs for Generative Design Exploration

Ariq Koh, Melvin Wong, Jiao Liu, Caishun Chen, Thiago de Jesus de Araujo Rios, Stefan Menzel, Yew Soon Ong

2026

Preprint:

Copyright ACM 2026. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in The Genetic and Evolutionary Computation Conference, https://doi.org/10.1145/3795101.3814651.

QD-LLMs: Quality-Diversity Optimization with LLMs for Generative Design Exploration

Ariq Koh*
Melvin Wong*†
Jiao Liu
College of Computing & Data Science
Nanyang Technological University
Singapore
ariq0001@ntu.edu.sg
wong1357@ntu.edu.sg
jiao.liu@ntu.edu.sg

Caishun Chen
Center for Frontier AI Research
Agency for Science, Technology and Research
Singapore
chen_caishun@a-star.edu.sg

Thiago Rios
Stefan Menzel
Honda Research Institute Europe
Offenbach am Main
Germany
thiago.rios@honda-ri.de
stefan.menzel@honda-ri.de

Yew Soon Ong
College of Computing & Data Science
Nanyang Technological University
Singapore
Center for Frontier AI Research
Agency for Science, Technology and Research
Singapore
asysong@ntu.edu.sg

The article has been accepted for publication in GECCO Companion ’26, San Jose, Costa Rica.

Copyright notice:

©Wong et al., ACM 2026. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in GECCO Companion ’26, San Jose, Costa Rica, http://dx.doi.org/10.1145/3795101.3814651.

QD-LLMs: Quality Diversity Optimization with LLMs for Generative Aerodynamic Design Search

Anonymous Authors

***Abstract*—Recent developments of text-to-3D generative models enable the synthesis of 3D geometries from free-form natural-language prompts. This presents an opportunity to systematically search for high-performing designs with diverse features across vast generative design spaces through such prompts. While the Quality-Diversity (QD) paradigm provides a principled framework for exploring an entire target feature space, there is a lack of QD variation operators that produce effective genotypes (or prompts) in free-form natural language, which are required to generate high-performing, domain-aligned phenotypes (or designs). Large Language Models (LLMs) alleviate this through learned semantic priors and conditional generation capabilities. To this end, we propose QD-Optimization with LLMs for Generative Aerodynamic Design Search (QD-LLMs), a novel QD framework that systematically explores a target feature space while ensuring domain validity of diverse designs through vision-language evaluations. QD-LLMs comprises an LLM allocator that selects elite LLM emitters to maximize search performance. The strategy of these specialized LLM emitters selects parent genotypes with high potential to yield high-performing, domain-aligned phenotypes, covering both explored and unexplored regions of the target feature space. These key components enable natural-language-coded variation operations that effectively explore the vast design spaces in text-to-3D generative models. Systematic experiments on aerodynamic vehicle design scenarios demonstrate that QD-LLMs substantially outperform both iterative zero-shot LLM prompting and real-coded QD methods, achieving 7.4× higher performance in generating valid, aerodynamically optimized designs than CMA-ME while maintaining comprehensive coverage of the target feature space.**

***Index Terms*—prompt evolution, engineering design optimization, quality-diversity optimization, large language models, text-to-3D, generative AI**

I. Introduction

Recent advances in text-to-3D generative models [1]–[3] have enabled the generation of 3D geometries from free-form natural language prompts, enabling exploration across vast generative design spaces. These models present a significant opportunity in engineering design to systematically search for diverse, high-performing solutions within a target feature space. In the early prototyping stage, engineers often explore diverse alternatives to understand how variations in physical properties may affect performance. Practitioners thus benefit from a collection of high-performing design candidates spanning a range of physical properties, motivating the need for systematic exploration tools that provide comprehensive coverage of a target design space. While text-to-3D generative models can produce diverse designs, translating desired physical targets into effective prompts is non-trivial. The relationship between prompt phrasing and resulting geometry is often unpredictable, making it infeasible to explore a target feature space through naive or brute-force prompting.

Quality-Diversity (QD) optimization offers a principled framework for structured exploration of a target feature space by distributing search pressure across the entire space rather than collapsing to a single optimum [4]. To drive exploration throughout the space, two main archival strategies are proposed in the literature that aim to partition the space into niches and maintain an archive of high-quality solutions within each niche. The first strategy implicitly partitions the space through clustering and prioritizes the exploration of new regions to promote diversity [5]. The other strategy explicitly decomposes the space into a grid map where each cell stores the best solution found for that region, unifying diversity exploration and fitness optimization [6]. This work investigates on the latter strategy. Furthermore, another notable QD component is the emitter, which selects parents from targeted niches to perform variation operations that encourage exploration of new regions in the feature space and improve the quality of existing solutions within niches [7]. Prior works introduce variation operators for various encodings to capture vital genetic information [4]. Among those, real-coded operators (such as Gaussian mutation [8] and CMA-ES [9]) are particularly prevalent in engineering design optimization due to its ability to exploit continuity in parameter spaces [10], [11]. Applying these real-coded operators to text-to-3D generative design search requires first encoding text prompts into continuous embeddings before conducting variation operations. Performing variations on these embeddings requires ensuring the perturbed embeddings remain coherent, preserving the semantic meaning and relationships between concepts. Otherwise, such variation operations will produce incoherent genotypes (or prompt offsprings), resulting in the synthesis of domain-invalid phenotypes (or designs) that may still satisfy the fitness objective and possess the desired features.

A viable alternative to this approach is to directly apply variation to free-form natural language prompts with large language models (LLMs). Recent work in literature demonstrates that LLMs are effective variation operators for natural-language prompts, ensuring that variations of textual prompts will always produce semantically coherent offspring through their learned priors [12], [13]. Furthermore, the conditional generation capabilities of LLMs enable offspring to be guided towards desired regions of a target feature space through in-context instructions, enabling directed exploration [13]. These capabilities open the possibility of designing natural-language-coded LLM emitters that leverage LLMs as variation operators

for text-conditioned generative search. We aim to investigate this possibility in this work.

To this end, we propose QD Optimization with LLMs for Generative Aerodynamic Design Search (QD-LLMs), a framework that systematically explores a target feature space in text-to-3D generative models while ensuring domain validity through vision-language evaluations. QD-LLMs offers two key components. First, three LLM emitters are introduced with different exploration strategies: (1) the LLM Uniform Emitter (QD-LLM/UE) addresses broad exploration by uniformly sampling genotypes across the archive, (2) the LLM Sparse Emitter (QD-LLM/SE) targets frontier expansion by biasing genotype selection towards the outer perimeter of occupied regions, and (3) the LLM Directional Emitter (QD-LLM/DE) performs targeted gap-filling in underexplored regions. Another key component of QD-LLMs is the LLM allocator, which coordinates a pool of LLM emitters to maximize search performance across the different phases of exploration. The allocator dynamically distributes computational budget across these LLM emitters based on archive state and performance history. Generated genotypes from LLM emitters are synthesized into 3D meshes and evaluated for both domain alignment and physical performance, enabling natural-language-coded variation operations that effectively explore vast generative design spaces. In summary, our contributions are as follows:

- This paper introduces QD Optimization with LLMs for Generative Aerodynamic Design Search (QD-LLMs), a novel framework that provides structured diversity exploration and fitness optimization in a target feature space within a text-to-3D generative model.
- Three LLM emitters (QD-LLM/UE, QD-LLM/SE, and QD-LLM/DE) with natural-language-coded variation operations are proposed to drive different exploration behaviors and produce coherent genotypes for text-conditioned generative search.
- An LLM allocator is introduced to maximize search performance by dynamically distributing computational budget across LLM emitters based on archive and performance statistics.
- Systematic evaluations of QD-LLMs on aerodynamic vehicle design scenario demonstrates 81% higher physical performance than zero-shot LLM baselines and 740% higher physical performance than latent-space QD methods in discovering high-performing, domain-valid designs while achieving systematic coverage of a predefined feature space.

II. Related Work

This work intersects with three research areas: text-to-3D generative models for design synthesis, Quality-Diversity optimization for structured exploration, and LLMs for evolutionary computation. This section reviews relevant work in each area and discusses recent efforts to combine LLMs with QD frameworks.

1) Text-to-3D Generative Models for Design: Text-to-3D models generate 3D geometry from natural language prompts. Shap-E [1] uses latent diffusion over implicit neural representations, completing generation in seconds per sample and making iterative design exploration feasible. Vision-language models such as BLIP-2 [14] enables automated evaluation of whether generated meshes match target domain descriptions. Prior work on generative design optimization has primarily operated in continuous latent spaces with domain-specific encoders [10], [11]. Text-to-3D models offer greater flexibility through natural language specification, though this introduces the challenge of optimizing in discrete prompt spaces.

2) Quality-Diversity Optimization: QD algorithms maintain archives of diverse, high-performing solutions rather than converging on a single optimum. MAP-Elites [6] partitions a behavior descriptor space into a grid and has been applied to robotics [15] and aerodynamic design [11]. Subsequent work improved variation operators. CMA-ME [7] introduced self-adaptive CMA-ES [9] emitters, while Vassiliades and Mouret [16] proposed directional variation (Iso+LineDD) exploiting correlations between archive elites. Recent surveys [4] categorize QD operators by encoding: Gaussian mutation [8] or CMA-ES for continuous spaces, swap and relocation for discrete spaces. Notably absent are operators for natural language encoding that consider semantic validity constraints.

3) Large Language Models in Evolutionary Algorithms: LLMs have emerged as an alternative to classical evolutionary operators. Liu *et al.* [12] proposes LLMs as zero-shot combinatorial optimizers, performing selection, crossover, and mutation via natural language prompts. In generative design, Rios *et al.* [17] demonstrates evolutionary optimization using text-to-3D for engineering applications, though without mechanisms for filtering impractical outputs. Wong *et al.* [18] addressed this by combining text-to-3D with vision-language models to penalize hallucinated designs. LLM-to-Phy3D [13] extends this with online iterative refinement using visual and physics-based feedback. These methods prioritize fitness optimization over diversity exploration within a target search space.

4) LLMs in Quality-Diversity Optimization: Two recent studies combine LLMs with QD optimization. QDAIF [19] addresses the lack of measurable quality and diversity metrics with LLM-based feedback, enabling QD applicability in qualitative domains where behavior descriptors are difficult to define manually. Gaier *et al.* [20] used QD to synthesize diverse datasets for fine-tuning language models in architectural design, but only for offline data generation rather than online optimization. Neither approach provides explicit behavioral targeting to steer generation toward specific underexplored archive regions. QD-LLMs addresses these gaps by coupling a grid-based QD archive with natural-language-coded variation operators that preserve semantic validity while supporting targeted exploration via in-context behavioral guidance.

III. QD-LLMs: Quality Diversity Optimization with LLMs for Generative Aerodynamic Design Search

This section presents QD-LLMs, a framework for structured exploration of text-conditioned generative design spaces, in the

context of an engineering design scenario. We formalize the problem, describe the framework architecture including the LLM allocator, and present the three natural-language-coded LLM emitters.

A. Problem Formulation

Let $\mathcal{Z}$ denote the space of natural language prompts and $\mathcal{X}$ the space of 3D designs/meshes. Given a target domain τ (e.g., "car"), each genotype (or prompt) follows the template $z =$ `"A` τ `in the shape of` ω`"`, where ω is the descriptive component that the algorithm optimizes. A text-to-3D generative model $G : \mathcal{Z} \to \mathcal{X}$ produces a phenotype (or design/mesh) $x = G(z)$ from a genotype z.

Following [18], we evaluate meshes using a fitness function that balances domain validity and physical performance:

$$f(x) = \lambda \cdot f_{\text{domain}}(x|\tau) + (1 - \lambda) \cdot f_{\text{physical}}(x), \tag{1}$$

where $\lambda \in [0, 1]$ is a weighting parameter, $f_{\text{domain}}(x|\tau) \in [0, 1]$ measures domain alignment with the target domain via vision-language models and $f_{\text{physical}}(x) \in [0, 1]$ quantifies physical performance on domain-specific metrics.

We maintain a grid-based archive $\mathcal{A}$ over a d-dimensional descriptor space $\mathcal{M} \subset \mathbb{R}^d$, where d is the number of predefined features of interest. A descriptor function $\mathbf{m} : \mathcal{X} \to \mathcal{M}$ maps each mesh to a d-dimensional feature vector $\mathbf{m}(x) = (m_1(x), \ldots, m_d(x))$, where each component function $m_i : \mathcal{X} \to \mathbb{R}$ extracts the i-th feature from a phenotype x. The descriptor space $\mathcal{M}$ is partitioned into a uniform grid with resolution $n_1 \times \cdots \times n_d$, yielding $\prod_{i=1}^{d} n_i$ cells. Each cell c represents a hyper-rectangular region in $\mathcal{M}$, and a mesh with descriptor $\mathbf{m}(x)$ is assigned to the cell whose region contains the point $\mathbf{m}(x)$. Each cell stores at most one prompt-mesh pair (z_c, x_c), preserving the highest fitness pair among all meshes mapped to that cell. The QD objective maximizes aggregate fitness across occupied cells:

$$\max_{\mathcal{A}} \sum_{c \in \mathcal{A}} f(x_c), \tag{2}$$

where x_c is the elite mesh in cell c with corresponding prompt z_c.

B. QD-LLMs Overview

QD-LLMs integrates grid-based Quality-Diversity archives with natural-language-coded variation operators for structured exploration of text-conditioned design spaces. Figure 1 illustrates the iterative process. Given the problem formulation in Section III-A, the framework operates as follows:

Initialization ($t = 1$): The archive $\mathcal{A}$ is seeded with N random genotypes, each denoted by $z \in \mathcal{Z}$. Each genotype z produces a phenotype $x = G(z)$, which is evaluated via $f(x)$ and assigned to cell c based on descriptor $\mathbf{m}(x)$. Genotype-phenotype pairs (z, x) are inserted into $\mathcal{A}$ following the elite preservation rule.

Iterative Search ($t = 2, \ldots, T$): Let $\mathcal{E}$ denote the set of available LLM emitters, where each LLM emitter produces one offspring per invocation. At each iteration t, the framework executes the following steps:

1. **LLM Emitter Selection:** At each iteration t, the LLM allocator selects N LLM emitters $\mathbf{e}^t = (e_1^t, \ldots, e_N^t)$ with $e_i^t \in \mathcal{E}$ based on archive and LLM emitter performance histories:
$$\mathbf{e}^t = \pi_{\text{LLM}}\left(\mathcal{H}_{\mathcal{A}}^t, \{\mathcal{H}_e^t\}_{e \in \mathcal{E}}, t, T, \mathcal{E}\right), \tag{3}$$
where π_{LLM} denotes an implicit LLM allocation policy based on its meta-prompt, $\mathcal{H}_{\mathcal{A}}^t$ and $\mathcal{H}_e^t$ are sliding window histories tracking recent archive statistics and LLM emitter e performance respectively, and T is the total number of iterations. The LLM reasons over these inputs before outputting $\mathbf{e}^t$.
2. **Genotypes Generation:** Each selected LLM emitter e_i^t executes its variation strategy, selecting parent genotypes from occupied cells in $\mathcal{A}^t$ and generating one offspring genotype z_o through natural-language-coded variation.
3. **Phenotypes Synthesis and Evaluations:** (MW: [Express this clearer mathematically.] Offspring genotypes are passed to the generative model G to produce phenotypes $x_o = G(z_o)$). Each phenotype is evaluated for domain alignment $f_{\text{domain}}(x_o|\tau)$ and physical performance $f_{\text{physical}}(x_o)$. The fitness $f(x_o)$ and descriptor $\mathbf{m}(x_o)$ are computed.
4. **Archive Update:** Each phenotype is mapped to cell c based on $\mathbf{m}(x_o)$. If c is unoccupied or $f(x_o) > f(x_c)$, the pair (z_o, x_o) is inserted into c, replacing any existing elite. The sliding window histories $\mathcal{H}_{\mathcal{A}}^t$ and $\{\mathcal{H}_e^t\}_{e \in \mathcal{E}}$ are updated based on insertion outcomes.

Termination: The process iterates until $t = T$, yielding a final archive $\mathcal{A}^T$ containing diverse, high-performing genotype and phenotype pairs across the descriptor space $\mathcal{M}$.

C. Evolutionary Operators in QD-LLMs

To fully unlock the capability of LLM-based operators to drive QD optimization, we design the LLm emitter set $\mathcal{E}$ comprising of the following LLM emitters which differ in parent selection and variation strategy, as illustrated in Fig. 2. Fig. 3 summarizes meta-prompts used.

1) Uniform LLM Emitter (QD-LLM/UE): QD-LLM/UE provides undirected exploration by selecting parent genotypes z_p uniformly at random from occupied cells in the archive $\mathcal{A}$. The LLM generates offspring genotypes z_o as creative variations without specified physical constraints, relying solely on its learned priors to produce diverse yet valid designs. This uniform selection ensures all regions of the archive contribute to variation.

2) Sparse LLM Emitter (QD-LLM/SE): QD-LLM/SE biases parent selection toward frontier regions to accelerate archive expansion. For each occupied cell c, the selection weight combines sparsity and curiosity:

$$w_c = (1 - \alpha) \cdot \text{sparsity}_c + \alpha \cdot \text{curiosity}_c, \tag{4}$$

where $\alpha \in [0, 1]$ is a balance parameter, $\text{sparsity}_c = 1/(1+n_c)$ with n_c the number of occupied neighbors, and curiosity_c

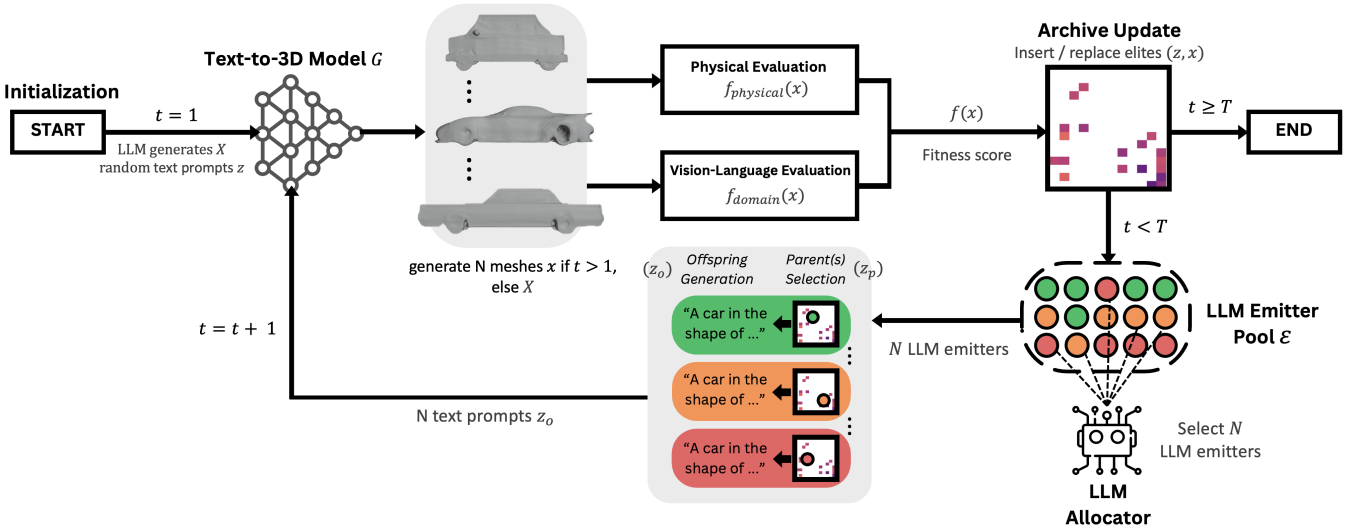

Fig. 1. QD-LLMs framework overview. Starting from initialization with X random prompts ($t = 1$), the system iteratively: (1) uses an LLM allocator to select N LLM emitters from the pool $\mathcal{E}$ based on archive and performance history, (2) generates offspring prompts z_o through LLM emitter operations on parent prompt(s) z_p, (3) synthesizes meshes x via text-to-3D model G and evaluates them for domain alignment and physical performance to compute $f(x)$, and (4) updates the grid-based archive with elite (z, x) pairs. The process repeats for T iterations.

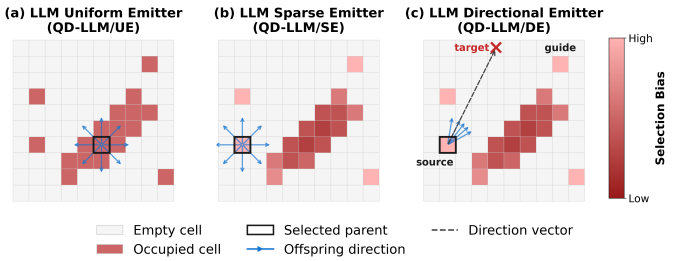

Fig. 2. LLM emitter strategies for parent selection and offspring generation. (a) QD-LLM/UE selects parents uniformly at random with unconstrained offspring directions. (b) QD-LLM/SE biases selection toward sparse regions with high curiosity scores. (c) QD-LLM/DE identifies a target cell, samples a source from sparse regions, and selects a guide elite along the source-to-target direction to extrapolate toward the target.

adapts based on offspring outcomes. The curiosity score [21] increases when offspring discover new cells or improve existing elites, decreases upon rejection, and decays by a factor each iteration to prevent stale scores from dominating. This mechanism balances exploitation of historically productive cells with exploration of geometrically sparse regions.

3) Directional LLM Emitter (QD-LLM/DE): QD-LLM/DE targets specific underexplored regions by conditioning generation on behavioral direction (Fig. 2c). The LLM emitter first computes a distance transform over the archive occupancy grid, assigning each empty cell a value equal to its Euclidean distance from the nearest occupied cell. The target cell c^* is the empty cell with maximum distance, representing the most underexplored region of $\mathcal{M}$. A source elite (z_s, x_s) is then sampled from sparse regions using the curiosity-weighted sparsity score w_c. To select a guide elite (z_g, x_g), the LLM emitter computes the direction vector from $\mathbf{m}(x_s)$ to c^* in normalized descriptor space and projects each occupied cell onto this vector. The guide is the elite with the largest positive projection, meaning the design that has moved furthest from the source toward the target. This source-guide pair establishes a behavioral trajectory that the LLM can extrapolate. The LLM receives the source genotype z_s, guide genotype z_g, and their behavioral deltas $\Delta\mathbf{m} = \mathbf{m}(x_g) - \mathbf{m}(x_s)$, with instructions to generate an offspring z_o that continues beyond the guide in the same direction. This enables the LLM to infer genotype-phenotype relationships from examples and generate genotypes that are likely to push designs towards the desired direction.

IV. Experiments

A. Experiment Groups

We evaluate QD-LLMs through two experiment groups: individual LLM emitters (CMA-ES, QD-LLM/UE, QD-LLM/SE, QD-LLM/DE) to isolate each variation operator, and multi-emitter coordination comparing QD-LLMs against CMA-ME [7]. Baselines include LLM Zero-Shot (iterative prompting with LLM feedback), CMA-ES (single emitter in prompt embedding space with grid archive), and CMA-ME (three CMA-ES emitters: improvement, random, optimizing).

B. Parameter Settings

We evaluate on aerodynamic vehicle design with target domain $\tau = \texttt{car}$. Shap-E serves as the generative model G, converting prompts $z \in \mathcal{Z}$ into 3D meshes $x \in \mathcal{X}$. Meshes are evaluated using a vision-language model [14] for domain alignment $f_{\text{domain}}(x|\tau)$ and a surrogate model [22] for drag estimation. The surrogate outputs $f_{\text{drag}}(x) \in [0, 1]$, and we define $f_{\text{physical}}(x) = 1 - f_{\text{drag}}(x)$. The LLM backbone is GPT-4o-mini with temperature sampled from $[0.9, 1.3]$. The fitness function $f(x)$ combines domain alignment and aerodynamic

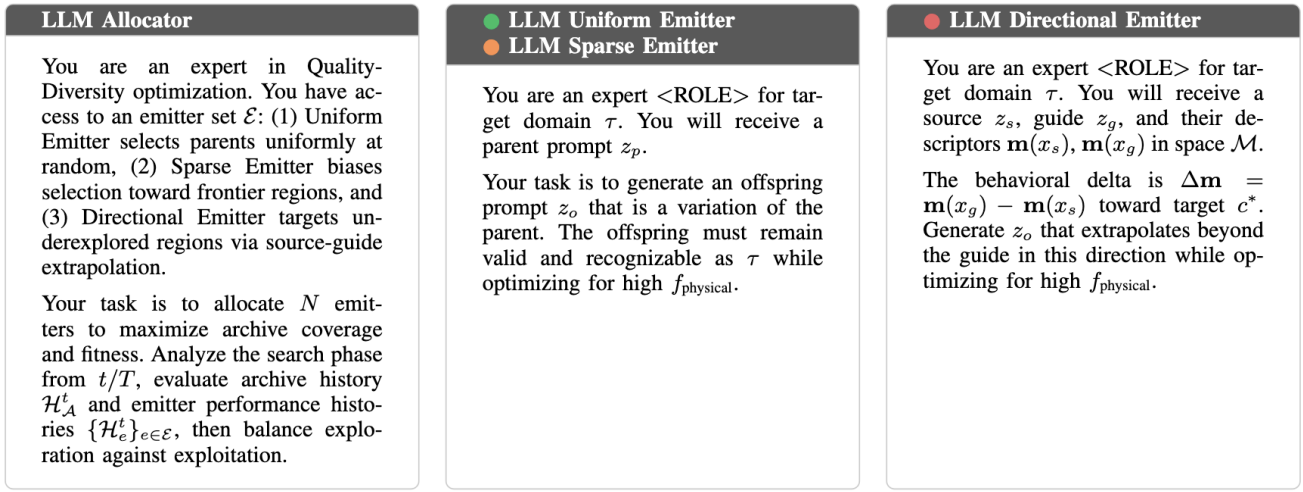

Fig. 3. Meta-prompts for LLM Allocator and Natural-Language-Coded LLM Emitters.

performance with $\lambda = 0.3$. The descriptor space $\mathcal{M} \subset \mathbb{R}^2$ is defined by volume and vertical center of gravity. The archive $\mathcal{A}$ discretizes $\mathcal{M}$ into a 20×20 grid, yielding 400 cells. For the curiosity-weighted sparsity score in QD-LLM/SE and QD-LLM/DE, we set $\alpha = 0.5$. Curiosity updates are $+1.0$ (new cell), $+0.5$ (improvement), -0.1 (rejection), with decay factor 0.95 per iteration.

All methods evaluate approximately 10,000 meshes for fair comparison. QD variants run for $T = 2000$ iterations with $N = 5$ offspring per iteration. LLM Zero-Shot uses 2000 iterations with population size 5 to match this budget. Archives are initially seeded with $X = 20$ random prompts. Results are averaged over 3 independent runs.

All experiments were conducted on a GPU compute node equipped with a single AMD EPYC 7713 CPU (16 of 64 cores utilized) and one NVIDIA A100-40GB GPU.

C. Evaluation Metrics

We adopt the Domain and Physical Alignment Rating (DPAR) [13], defined as the ratio of domain alignment to drag coefficient. DPAR penalizes methods that achieve low drag through semantically invalid designs, capturing the balance between physical performance and domain validity central to text-conditioned generative design. Let $\mathcal{A}$ denote the archive, X_t the solutions generated at iteration t, and $X = \bigcup_t X_t$ all generated solutions.

Averaged metrics summarize per-iteration performance:

- **Average Domain Score**: $\frac{1}{T}\sum_{t=1}^{T}\frac{1}{|X_t|}\sum_{x\in X_t} f_{\text{domain}}(x)$ — mean domain score per solution, per iteration.
- **Average QD-Score**: $\frac{1}{T}\sum_{t=1}^{T}\frac{1}{|X_t|}\sum_{x\in X_t}\frac{f_{\text{domain}}(x)}{f_{\text{drag}}(x)}$ — mean domain/drag ratio per solution, per iteration.
- **Average DPAR**: $\frac{1}{|X|}\sum_{x\in X}\frac{f_{\text{domain}}(x)}{f_{\text{drag}}(x)}$ — mean domain/drag ratio across all generated solutions.

Cumulative metrics plot QD archive growth over iterations:

- **Coverage**: $\sum_{x\in\mathcal{A}} f_{\text{domain}}(x)$ — total domain score in the archive.
- **QD-Score**: $\sum_{x\in\mathcal{A}}\frac{f_{\text{domain}}(x)}{f_{\text{drag}}(x)}$ — total quality-diversity score in the archive.
- **DPAR**: $\frac{1}{|\mathcal{A}|}\sum_{x\in\mathcal{A}}\frac{f_{\text{domain}}(x)}{f_{\text{drag}}(x)}$ — mean quality-diversity across the archive.

Averaged metrics capture *how well the method generates valid designs*; cumulative metrics capture *what the archive contains at each iteration*.

V. Results and Discussion

Tables I and II summarize performance across all evaluated methods. Figure 5 presents the evolution of coverage, QD-score, and DPAR over 2000 iterations. Figures 4, 6, and 7 provide archive visualizations for LLM Zero-Shot, individual LLM emitters and multi-emitter coordination respectively. Figure 8 illustrates emitter allocation patterns across search phases.

Structured diversity with LLM variation. We first examine whether combining structured diversity maintenance with LLM variation is benefits design exploration. Three methods represent different component combinations: LLM Zero-Shot (LLM variation without archive structure), CMA-ES (archive structure without LLM variation), and QD-LLM/UE (both components). LLM Zero-Shot achieves a DPAR of 0.514 (Table I). The model's learned priors constrain outputs to valid vehicle descriptions, producing recognizable designs across runs (Fig. 4). However, the absence of archive structure provides no mechanism for systematic behavioral coverage or preservation of high-performing solutions across behavioral niches. Diversity emerges incidentally from sampling stochasticity rather than explicit selection pressure.

CMA-ES achieves a DPAR of only 0.015 (Table I). Figure 6 reveals that while the archive spans the feature space, outputs consist of amorphous forms, fragmented geometry, and abstract shapes that achieve low drag through trivial solutions such as flat discs. The archive exhibits physical diversity but lacks domain validity.

TABLE I
INDIVIDUAL LLM EMITTER PERFORMANCE (AVERAGED ACROSS ALL ITERATIONS)

Metric	QD-LLM/DE	QD-LLM/SE	QD-LLM/UE	CMA-ES	LLM Zero-Shot
QD-Score (Avg)	**0.9228 ± 0.0679**	0.8003 ± 0.0249	0.8088 ± 0.0289	0.0110 ± 0.0119	0.5158 ± 0.0024
Domain Score (Avg)	**0.8870 ± 0.0688**	0.7588 ± 0.0140	0.7580 ± 0.0232	0.0076 ± 0.0056	0.6248 ± 0.0047
DPAR (Avg)	**0.9296 ± 0.0639**	0.8161 ± 0.0251	0.8232 ± 0.0280	0.0147 ± 0.0114	0.5137 ± 0.0023

TABLE II
MULTI-EMITTER PERFORMANCE (AVERAGED ACROSS ALL ITERATIONS)

Metric	QD-LLMs	CMA-ME
QD-Score (Avg)	**0.8828 ± 0.0208**	0.0956 ± 0.0462
Domain Score (Avg)	**0.9771 ± 0.0280**	0.0711 ± 0.0319
DPAR (Avg)	**0.9200 ± 0.0246**	0.1238 ± 0.0531

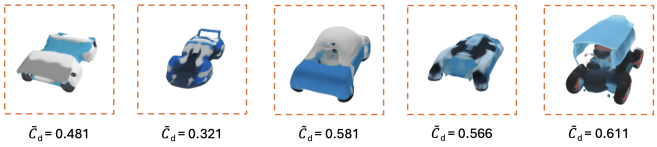

Fig. 4. Final population from LLM Zero-Shot. Valid vehicle designs are generated with visible variation, but without archive structure, diversity emerges from sampling stochasticity rather than selection pressure toward systematic behavioral coverage.

QD-LLM/UE achieves substantially higher domain alignment (0.758 vs 0.008) and DPAR (0.823 vs 0.015). Figure 6 shows that QD-LLM/UE generated designs maintain vehicle appearance while varying systematically across the descriptor dimensions.

The 56× DPAR improvement reflects a fundamental limitation of latent space mutation in text conditioned domains: small perturbations in prompt embeddings produce incoherent outputs and domain invalid designs. Natural-language-coded variation circumvents this by operating directly in prompt space, where learned priors ensure semantic coherence. Neither zero shot prompting (lacks systematic coverage) nor archive based search without LLM variation (produces invalid designs) suffices alone.

Targeted versus undirected variation. Having established the advantages of natural-language-coded variation, we examine whether explicitly directing generation toward underexplored regions improves exploration efficiency. QD-LLM/UE, QD-LLM/SE, and QD-LLM/DE represent increasing levels of feature space targeting: uniform parent selection (QD-LLM/UE), sparsity-biased selection (QD-LLM/SE), and directional variation with in-context guidance (QD-LLM/DE).

QD-LLM/SE achieves comparable aggregate performance to QD-LLM/UE in average domain alignment (0.759 vs 0.758) and DPAR (0.816 vs 0.823). At the same time, fig. 6 shows that QD-LLM/SE expands into frontier regions inaccessible to SUE, indicating complementary coverage patterns rather than strict metric dominance. QD-LLM/DE achieves 17% higher average domain alignment than QD-LLM/UE (0.887 vs 0.758) and 14% higher QD-score (0.923 vs 0.809) as shown in Table I. Figure 5 supports these findings with steeper coverage and QD-score trajectories. Archive visualizations in fig. 6 indicate that QD-LLM/DE fills boundary regions that QD-LLM/UE leaves sparse. This improvement demonstrates that LLMs can infer prompt-feature relationships from in-context examples without explicit training on the target domain, suggesting a generalizable principle for LLM-QD integration: target objectives may be communicated more effectively through demonstrations than through explicit instructions.

Adaptive LLM emitter allocation. We evaluate whether LLM-based coordination of multiple LLM emitters outperforms single LLM emitter strategies by comparing QD-LLMs against QD-LLM/DE and CMA-ME.

QD-LLMs achieves the highest average domain alignment across all evaluated methods (Table II). Relative to QD-LLM/DE, QD-LLMs improves average domain alignment by 10% (0.977 vs 0.887). Relative to CMA-ME, improvements are substantially larger, with 13.7× higher average domain alignment (0.977 vs 0.071), 9.2× higher QD-score (0.883 vs 0.096), and 7.4× higher DPAR (0.920 vs 0.124). Fig. 5 confirms that QD-LLMs achieves the highest final coverage, QD-score, and DPAR. Archive visualization (Fig. 7) show that QD-LLMs fills regions that individual LLM emitters leave sparse.

Fig. 8 reveals emergent phase-dependent allocation trends. Early search favors QD-LLM/UE for broad archive seeding. As coverage increases, allocation shifts toward QD-LLM/DE for systematic gap-filling. Late search balances QD-LLM/DE and QD-LLM/SE for targeting sparse regions and frontier expansion. Unlike bandit-based allocation which responds to scalar reward signals, the LLM allocator dynamically bases its allocation decision on why particular LLM emitters may be appropriate given the current search state. The three LLM emitters serve complementary functional roles. QD-LLM/UE provides broad exploration across the feature space. QD-LLM/SE concentrates on frontier expansion. QD-LLM/DE targets specific underexplored regions. Individual LLM emitters each leave characteristic gaps, visible as dark regions in domain heatmaps (Fig. 6). QD-LLMs manages to cover these gaps through dynamic strategy switching based on archive state (Fig. 7). These results suggest that LLMs can function as flexible meta-controllers for optimization, adapting strategy through contextual reasoning rather than fixed allocation rules.

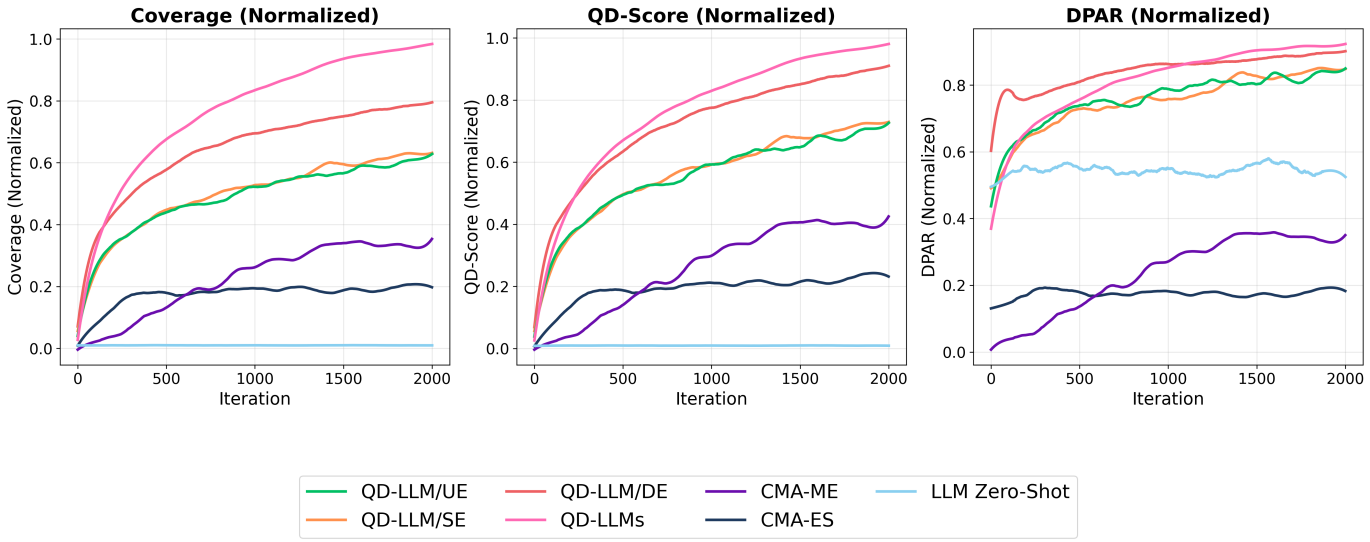

Fig. 5. Evolution of Coverage, QD-Score, and DPAR over 2000 iterations. Coverage (left) measures the sum of domain scores across occupied archive cells, reflecting both cell occupancy and design validity. QD-Score (center) combines domain alignment with aerodynamic fitness. DPAR (right) captures the ratio of domain alignment to drag coefficient, penalizing methods that achieve low drag through semantically invalid designs. LLM based methods exhibit steady improvement throughout search, with QD-LLMs achieving the highest final coverage and QD-Score. Latent space methods (CMA-ES, CMA-ME) plateau early as invalid designs accumulate without meaningful improvement. LLM Zero Shot achieves moderate DPAR but lacks systematic coverage improvement.

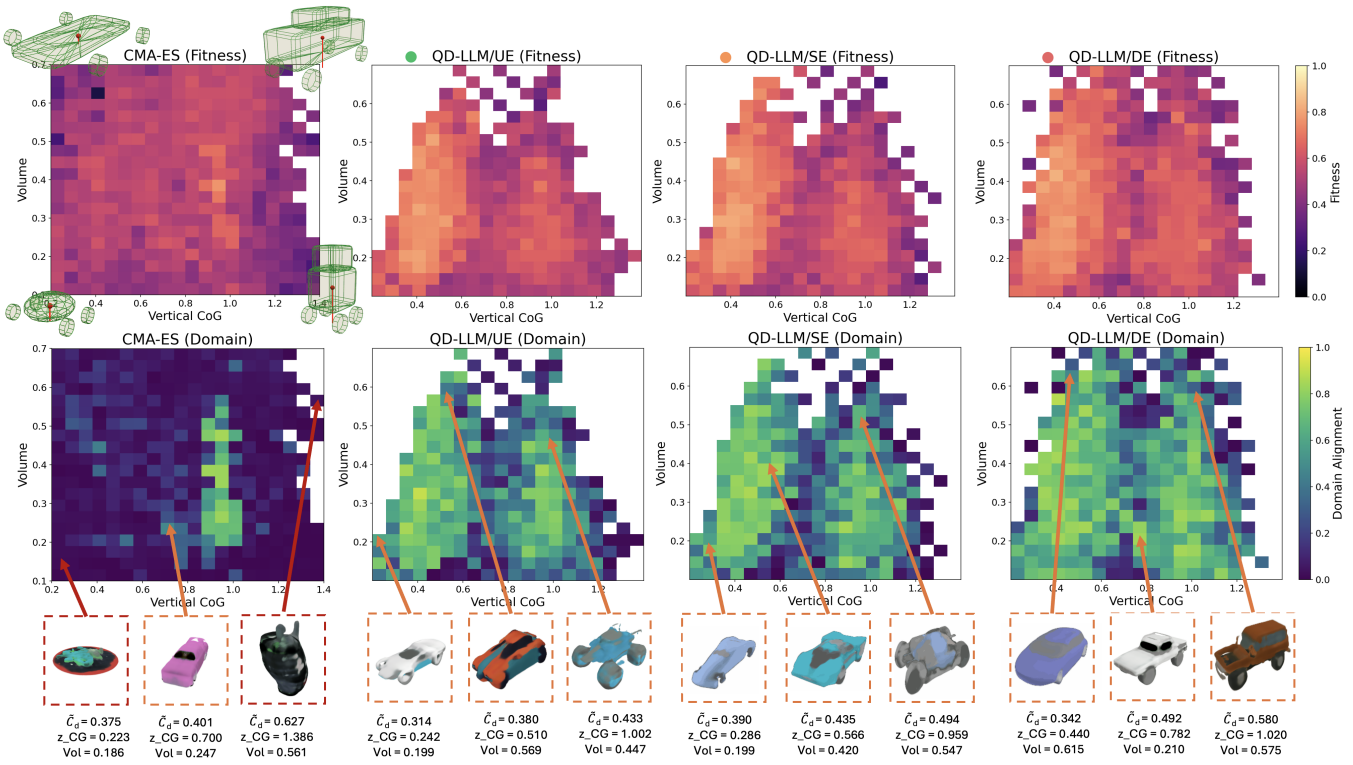

Fig. 6. Archive visualization for individual LLM emitters: CMA-ES, QD-LLM/UE, QD-LLM/SE, and QD-LLM/DE (left to right). Top row shows objective scores and bottom row shows domain alignment. Valid designs are outlined in orange and invalid designs in red. Wireframe examples at the corners illustrate reference vehicle shapes at behavioral extremes with vertical center of gravity increasing along the horizontal axis and volume along the vertical axis. CMA-ES fills cells across the behavioral space but produces semantically invalid designs (amorphous blobs, fragmented geometry) that achieve low drag through trivial solutions rather than aerodynamic shaping, resulting in near zero domain alignment despite high occupancy. Among LLM emitters, QD-LLM/UE achieves broad interior coverage but leaves frontier regions sparse due to undirected variation. QD-LLM/SE expands into these frontiers through sparsity biased parent selection. QD-LLM/DE achieves greatest coverage by systematically targeting underexplored regions via in context behavioral guidance using source guide pairs with measure deltas.

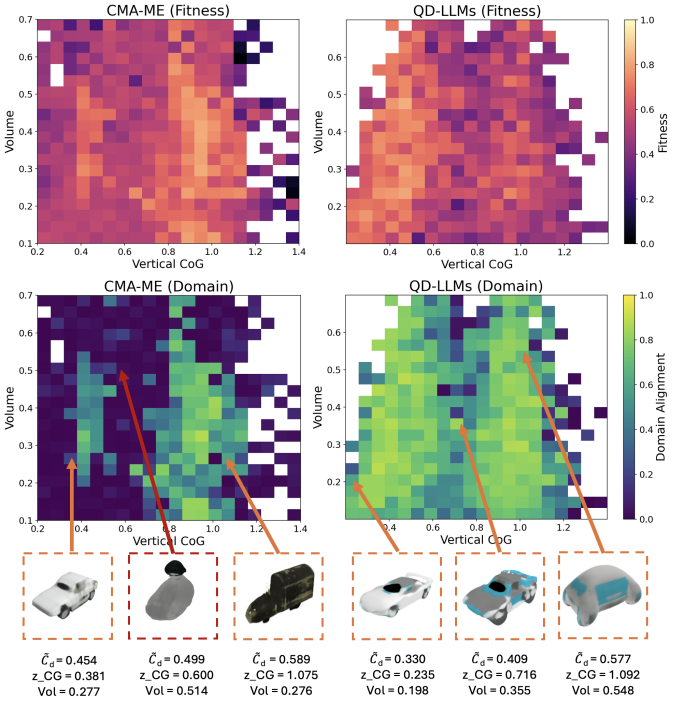

Fig. 7. Archive visualization for multi-emitter methods. Valid designs are outlined in orange; invalid designs in red. CMA-ME (left) fills cells with semantically invalid designs. QD-LLMs (right) achieves high coverage with consistent domain validity.

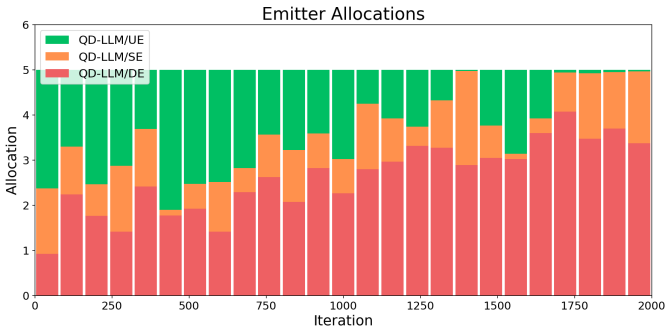

Fig. 8. LLM emitter allocation patterns across search phases. LLM-UE dominates early search; LLM-DE allocation increases in late search; LLM-SE maintains consistent allocation throughout.

VI. Conclusion

This paper presented QD-LLMs, a framework integrating grid-based QD archives with three natural-language-coded LLM emitters for structured exploration of text-conditioned generative design spaces. QD-LLM/UE for broad exploration, QD-LLM/SE for frontier expansion, and QD-LLM/DE for targeted exploration. An LLM allocator coordinates LLM emitter selection via reasoning over search dynamics.

Experimental results on aerodynamic vehicle design validate the effectiveness of QD-LLMs, achieving 81% higher DPAR than zero-shot LLM prompting and 740% higher DPAR than CMA-ME. The framework inherits limitations from its underlying models, particularly Shap-E's generation quality and the LLM's domain coverage. Nevertheless, the approach generalizes to any domain where text-to-X generative models exist and semantic validity is required, providing systematic exploration with principled diversity maintenance. Future work includes quantifying directional variation reliability and comparing the LLM allocator against bandit-based allocation.

References

[1] H. Jun and A. Nichol, "Shap-e: Generating conditional 3d implicit functions," arXiv preprint arXiv:2305.02463, 2023.

[2] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen, "Point-e: A system for generating 3d point clouds from complex prompts," arXiv preprint arXiv:2212.08751, 2022.

[3] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," arXiv preprint arXiv:2209.14988, 2022.

[4] H. Qin, Y. Xiang, H. Zhang, Y. Han, Y. Wang, X. Tao, and Y. Liu, "A survey on quality-diversity optimization: Approaches, applications, and challenges," Swarm and Evolutionary Computation, vol. 100, p. 102240, 2026.

[5] J. Lehman and K. O. Stanley, "Evolving a diversity of virtual creatures through novelty search and local competition," in Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation (GECCO '11). New York, NY, USA: Association for Computing Machinery, 2011, pp. 211–218.

[6] J.-B. Mouret and J. Clune, "Illuminating search spaces by mapping elites," arXiv preprint arXiv:1504.04909, 2015.

[7] M. C. Fontaine, J. Togelius, S. Nikolaidis, and A. K. Hoover, "Covariance matrix adaptation for the rapid illumination of behavior space," in Proc. of GECCO, 2020, pp. 94–102.

[8] K. Chatzilygeroudis, A. Cully, V. Vassiliades, and J.-B. Mouret, "Quality-diversity optimization: A novel branch of stochastic optimization," arXiv preprint arXiv:2012.04322, 2020.

[9] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," Evolutionary Computation, vol. 9, no. 2, pp. 159–195, 2001.

[10] W. Chen, K. Chiu, and M. Fuge, "Airfoil design parameterization and optimization using bézier generative adversarial networks," arXiv preprint arXiv:2006.12496, 2020.

[11] A. Gaier, A. Asteroth, and J.-B. Mouret, "Data-efficient design exploration through surrogate-assisted illumination," arXiv preprint arXiv:1806.05865, 2018.

[12] S. Liu, C. Chen, X. Qu, K. Tang, and Y.-S. Ong, "Large language models as evolutionary optimizers," in Proceedings of the 2024 IEEE Congress on Evolutionary Computation (CEC), 2024, pp. 1–8.

[13] M. Wong, Y. Lyu, T. Rios, S. Menzel, and Y.-S. Ong, "Llm-to-phy3d: Physically conform online 3d object generation with llms," arXiv preprint arXiv:2506.11148, 2025.

[14] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," arXiv preprint arXiv:2301.12597, 2023.

[15] A. Cully, J. Clune, D. Tarapore, and J.-B. Mouret, "Robots that can adapt like animals," Nature, vol. 521, no. 7553, pp. 503–507, 2015.

[16] V. Vassiliades and J.-B. Mouret, "Discovering the elite hypervolume by leveraging interspecies correlation," in Proc. of the Genetic and Evolutionary Computation Conference (GECCO '18). ACM, 2018, pp. 149–156.

[17] T. Rios, S. Menzel, and B. Sendhoff, "Large language and text-to-3d models for engineering design optimization," arXiv preprint arXiv:2307.01230, 2023.

[18] M. Wong, T. Rios, S. Menzel, and Y.-S. Ong, "Prompt evolutionary design optimization with generative shape and vision-language models," in 2024 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2024, pp. 1–8.

[19] H. Bradley, A. Dai, H. Teufel, J. Zhang, K. Oostermeijer, M. Bellagente, J. Clune, K. Stanley, G. Schott, and J. Lehman, "Quality-diversity through ai feedback," arXiv preprint arXiv:2310.13032, 2023.

[20] A. Gaier, J. Stoddart, L. Villaggi, and S. Sudhakaran, "Generative design through quality-diversity data synthesis and language models," in Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '24), 2024, pp. 823–831.

[21] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in International Conference on Machine Learning. PMLR, 2017, pp. 2778–2787.

[22] B. Song, C. Yuan, F. Permenter, N. Arechiga, and F. Ahmed, "Surrogate modeling of car drag coefficient with depth and normal renderings," arXiv preprint arXiv:2306.06110, 2023.