IMPERIAL

MSc Individual Project

Imperial College London

Department of Computing

Diversify Guided 3D Generation via Repulsive 3D Gaussian Splatting

Author:
Sijeong Kim

Supervisor:
Dr. Li, Yingzhen

Second Marker:
Dr. Deng, Jiankang

Submitted as the Final Report for the MSc Individual Project in Computing (Artificial Intelligence and Machine Learning) at Imperial College London

September 24, 2025

Abstract

Text-to-3D generation has recently attracted significant attention for its ability to create 3D assets directly from natural language prompts. A dominant approach is *Score Distillation Sampling (SDS)*, which leverages pretrained 2D text-to-image diffusion models to guide optimisation of 3D representations. While effective, SDS suffers from well-documented limitations: *mode collapse*, where runs converge to near-identical shapes, and *geometric inconsistencies*, such as the Janus problem. These issues restrict fidelity and diversity, limiting the scalability of SDS-based pipelines in practice.

This dissertation introduces lightweight, modular strategies to address these limitations by incorporating **feature-space repulsion** into the DreamGaussian pipeline. Two complementary formulations—*Repulsive Latent Score Distillation (RLSD)* and *Stein Variational Gradient Descent (SVGD)*— are lifted from latent space into *semantic feature space* (CLIP/DINO embeddings of renders). This encourages parallel 3D reconstructions (particles) to spread apart in semantically meaningful directions, counteracting collapse while retaining SDS's fidelity.

A **comprehensive evaluation framework** is developed, combining fidelity, diversity, and cross-view consistency into a unified protocol, alongside ablations of kernel choice, repulsion strength, guidance scale, and kernel temperature. Metrics are complemented by representation-level PCA analyses and a user study.

Results show that the proposed RLSD–RBF configuration **increases semantic diversity from 0.132 to 0.262** (Δ+0.130, $\sim$98% relative gain), while **fidelity remains effectively constant** (0.391 vs. 0.397, $\Delta-0.006$). Cross-view consistency decreases slightly ($0.853 \rightarrow 0.828$, $\Delta-0.025$) but remains high ($\mathcal{C}>0.83$). In human evaluation (n=41), participants rated our method as substantially more diverse (up to +2.05 Likert points, $p < 0.001$), while realism preferences were balanced and not statistically distinguishable.

In summary, this work establishes feature-space repulsion as an efficient and scalable strategy for enhancing semantic variety in text-to-3D generation, addressing core limitations of SDS while preserving both fidelity and coherence.

Acknowledgments

I am deeply grateful to my supervisor, Dr. Yingzhen Li, for unwavering guidance and support throughout this project. I owe particular thanks to PhD candidate Zhengrui Xiang for his steady assistance and thoughtful advice, and to Alex Pondaven and Harrison for their helpful input at the early stages. I also thank all members of our lab for fostering a collaborative workspace that made my time there truly enjoyable. I am especially thankful to my friends for believing in me, caring for me, and generously giving their time. Above all, I am profoundly grateful to my family for their endless love and encouragement, which sustained me with strength and joy. Lastly, I would like to acknowledge my own perseverance; irrespective of any objective assessment of success, I am grateful for the effort I invested over this past year abroad.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Objectives

Text-to-3D generation is the task of creating 3D digital assets directly from natural language descriptions. This technology has the potential to transform domains such as gaming, robotics, virtual reality, and digital content creation, by dramatically reducing the manual effort and specialised expertise traditionally required for 3D modelling. Instead of hand-crafting meshes or textures, users can describe objects or scenes in plain language and receive detailed 3D models in return. In this way, text-to-3D generation democratises access to 3D content creation and accelerates design iteration [7, 8, 9, 10].

At the heart of these advances lie diffusion models. These generative models synthesise data by starting from random noise and iteratively refining it through a learned denoising process. In 2D image generation, diffusion models [11] have surpassed previous approaches such as GANs [12], delivering state-of-the-art results in synthesis, inpainting, and super-resolution. Their stability, interpretability, and ability to produce coherent and diverse samples make them a powerful foundation for downstream applications. The success of diffusion in 2D has naturally motivated its extension to 3D domains, despite the added complexity.

1.2 Challenges

Extending diffusion to 3D presents several fundamental challenges. Unlike 2D images, 3D data is inherently more complex, spanning unstructured and unordered representations such as meshes, point clouds, or implicit functions—each with trade-offs in memory usage, resolution, and rendering efficiency [13, 2, 14, 15]. The scarcity of large, high-quality 3D datasets further limits the direct training of diffusion-based 3D generative models, making it difficult to achieve the same level of fidelity and diversity seen in 2D applications.

A widely adopted workaround is to leverage pretrained 2D diffusion models as priors to guide the optimisation of 3D representations. Score Distillation Sampling (SDS) [7] has emerged as the canonical method for this purpose. While SDS enables zero-shot text-to-3D generation without paired supervision, it exhibits significant shortcomings. Most notably, SDS often produces *over-smoothed and low-diversity outputs*, with multiple runs collapsing to nearly identical shapes [7, 8, 6]. This *mode collapse* severely restricts the ability to explore diverse design variations. By contrast, geometric inconsistency across views (e.g., the Janus problem) is a broader limitation of 2D-guided 3D generation but lies outside the scope of this dissertation [16, 17].

Figure 1.1 illustrates this mode collapse problem, showing DreamGaussian [1] generations for the prompt “a photo of a hamburger” at step 500. Despite stochastic initialisation across eight random seeds, all runs converge to nearly identical outputs. Such lack of diversity is not only a technical pathology but also a practical bottleneck: in creative domains such as gaming, virtual reality, and digital content production, generating near-identical assets across runs undermines the exploratory and creative potential of text-to-3D systems.

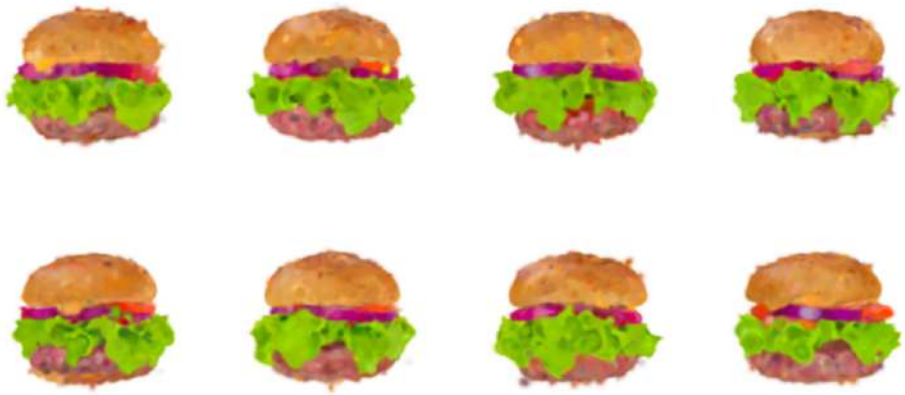


Figure 1.1: Illustration of mode collapse in DreamGaussian [1]. Generations for the prompt "a photo of a hamburger" at step 500 across multiple random seeds converge to almost identical shapes, highlighting the lack of diversity in vanilla SDS guidance.

1.3 Contributions

This dissertation aims to address the limitations of current SDS-based text-to-3D pipelines, with a particular focus on mitigating mode collapse while preserving fidelity and efficiency. To this end, we investigate lightweight, diversity-enhancing mechanisms integrated into an efficient 3D backbone.

Concretely, the contributions are as follows:

- **Mitigating mode collapse:** We incorporate kernel-based repulsive mechanisms [6, 18] that encourage parallel particles to explore semantically distinct yet text-aligned solutions.
- **Adopting 3D Gaussian Splatting (3DGS):** We employ 3DGS [2] as the core representation, leveraging its efficient and differentiable rendering to support scalable, stable optimisation.
- **Comprehensive evaluation:** We systematically evaluate the proposed enhancements in terms of diversity, fidelity, and computational efficiency, combining quantitative metrics with qualitative human preference studies.

By integrating repulsive guidance with an efficient rendering backbone, this dissertation contributes a practical, scalable, and more diverse pipeline for text-driven 3D asset generation. The resulting system not only holds potential for creative industries but also provides a foundation for future research in 3D generative modelling.

Chapter 2

Literature Review

Roadmap. We begin by contrasting implicit (NeRF) and explicit (3DGS) 3D representations. Next, we review diffusion-guided optimisation via Score Distillation Sampling (SDS) and its limitations. We then survey subsequent pipelines and classify how they introduce (or fail to introduce) diversity, highlighting distributional versus kernel-based approaches. This trajectory reveals a gap: existing repulsion methods act only in diffusion *latents* or remain unexplored; we instead study repulsion in *semantic features* atop 3DGS.

2.1 3D Representations for Generative Modelling

2.1.1 NeRF vs 3DGS

Neural Radiance Fields (NeRF) [13] represent a 3D scene as a continuous volumetric function parameterised by a neural network. For each sampled 3D point and viewing direction, NeRF predicts both density and view-dependent radiance, which are integrated along rays via differentiable volume rendering. This formulation enables highly realistic novel-view synthesis from a sparse set of input images, and NeRF quickly became a standard backbone for generative pipelines due to its photorealism and differentiability.

However, NeRF suffers from high computational cost: training requires tens of hours on a single scene due to dense ray sampling, and rendering is slow because radiance must be integrated along thousands of rays. These limitations make NeRF difficult to scale for generative text-to-3D pipelines that require optimisation from scratch for each prompt. Furthermore, NeRF's implicit representation often leads to overfitting or floating artefacts when supervision is weak, a common issue in SDS-guided pipelines. Early implicit representations such as Occupancy Networks [14] and DeepSDF [15] similarly highlight the strengths and weaknesses of implicit neural 3D encodings.

In contrast, 3D Gaussian Splatting (3DGS) [2] represents scenes as explicit anisotropic Gaussians projected into screen space. This enables real-time rasterisation and efficient gradient flow, drastically reducing training time. DreamGaussian [1] demonstrated that substituting NeRF with 3DGS not only improves efficiency but also yields more stable optimisation, making 3DGS the preferred backbone for modern text-to-3D pipelines.

2.2 Text-to-3D via Diffusion Guidance

2.2.1 Diffusion Models as 2D Priors

Denoising Diffusion Probabilistic Models (DDPMs) [11] are a class of generative models that learn to reverse a Markovian noising process. Starting from pure Gaussian noise, a diffusion model iteratively denoises the latent variable $\mathbf{x}_t$ using a learned noise predictor $\epsilon_\theta(\mathbf{x}_t, t, y)$ conditioned on a text prompt y. This iterative denoising process yields high-quality, text-aligned images when trained on large-scale datasets.

Recent works such as Stable Diffusion [19], Imagen [20], and DALLE-2 [21] demonstrated that diffusion models trained on large-scale datasets can produce high-quality, text-aligned images. They demonstrated

that diffusion models can be trained in a latent space rather than pixel space, drastically reducing computation while retaining generative fidelity. As a result, pretrained 2D diffusion models have become a powerful prior for downstream generative tasks, including text-to-3D.

2.2.2 Score Distillation Sampling (SDS)

Score Distillation Sampling (SDS) [7] is a foundational technique that enables optimisation of a 3D representation using only a pretrained 2D diffusion prior. Given a noisy rendered image $\mathbf{x}_t$ from a differentiable 3D scene parameterisation and a text prompt y, SDS minimises the discrepancy between the diffusion model's predicted noise $\epsilon_\theta(\mathbf{x}_t, t, y)$ and the sampled noise ϵ:

$$\mathcal{L}_{\text{SDS}} = \left\|\epsilon_\theta(\mathbf{x}_t, t, y) - \epsilon\right\|^2. \tag{2.1}$$

This loss can be interpreted as a gradient distillation process: the frozen diffusion model acts as a teacher, providing semantic gradients that guide the 3D parameters towards text-consistent renderings. SDS is simple and data-efficient, requiring no 3D supervision. However, it introduces several pathologies: Known issues include mode collapse, Janus artefacts, and cross-view inconsistency (reported across SDS-based pipelines [7, 9, 1]). Importantly, the SDS formulation and its reliance on latent-variable encoding/decoding are directly grounded in the Stable Diffusion framework [19].

Mode collapse in SDS. Mode collapse has been repeatedly documented in SDS-based pipelines. DreamFusion [7] explicitly reports that SDS optimisation often leads to "over-smoothed" and low-diversity shapes, with generations collapsing to similar forms despite different random seeds. Score Jacobian Chaining (SJC) [22]—a concurrent work to DreamFusion—acknowledges analogous limitations, reinforcing that the issue is structural to SDS guidance rather than implementation-specific. ProlificDreamer [8] provides a more systematic analysis, showing that vanilla SDS tends to produce near-identical outputs across seeds, particularly for common categories. They further propose distributional updates (VSD) as a mitigation strategy, highlighting the need for improved guidance mechanisms.

2.2.3 DreamFusion and Successors

DreamFusion [7] pioneered the SDS pipeline with NeRF [13] as the 3D backbone. While it established the feasibility of text-to-3D generation without 3D supervision, it suffered from low resolution, expensive optimisation, and geometric instability.

Magic3D [9] improved resolution via a two-stage optimisation, but as it still relied on SDS, it inherited the same pathologies such as mode collapse and cross-view inconsistency.

DreamGaussian [1] introduced a more radical change by replacing NeRF with 3D Gaussian Splatting (3DGS) [2], enabling real-time rasterisation, faster convergence, and greater training stability. Despite these advantages, DreamGaussian continues to inherit the limitations of SDS. Given its efficiency and stability, we adopt DreamGaussian as our primary baseline, allowing us to directly assess the impact of introducing kernel-based repulsion mechanisms on top of a strong 3DGS-based pipeline.

2.3 Probabilistic and Repulsive Extensions

Where diversity acts. We categorise methods by (i) *3D representation* and *guidance* and, crucially, (ii) how/where they introduce diversity. *Kernel repulsion* denotes similarity-kernel forces that separate parallel particles (RLSD uses a log-sum-exp kernel field; SVGD uses a Stein field that combines score attraction with a repulsive kernel term). *Acts in* specifies the space where this pressure is applied: *Latent* refers to the diffusion latent z; *Feature* refers to a semantic image embedding (e.g., CLIP/DINO); and *Distribution* indicates distributional updates (e.g., VSD under Wasserstein gradient flows, "WGF") that operate at the level of the scene distribution rather than pairwise repulsion. A side-by-side summary is given in Table 2.1.

As summarised in Table 2.1, existing text-to-3D pipelines differ not only in their underlying 3D representation and guidance mechanism, but also in how (or whether) they introduce diversity. This highlights the structural limitations of SDS-based approaches, and points towards the potential of feature-space repulsion, which we return to in Sec. 2.4.3.

Table 2.1: Text-to-3D pipelines: *how/where* diversity is introduced, with key takeaways.

Method	3D Rep.	Guid.	Diversity Mech.	Acts in	Pros and Cons
DreamFusion [7]	NeRF	SDS	N/A	N/A	+ Canonical SDS baseline; – collapse; cross-view gaps
Magic3D [9]	NeRF	SDS	N/A	N/A	+ Higher-res; – inherits SDS collapse
DreamGaussian [1]	**3DGS**	SDS	N/A	N/A	+ Efficient, stable; – collapse persists under SDS
ProlificDreamer [8]	NeRF	VSD	WGF (distributional)	Distribution	+ Principled diversity; – higher compute; often prior adaptation (e.g., LoRA)
RLSD [6]	*method-agnostic*	SDS	Kernel repulsion (log-sum-exp)	Latent	+ Plug-and-play; – kernel/bandwidth sensitive
Ours	**3DGS**	SDS	Kernel repulsion (SVGD/RLSD)	Feature	+ Feature-space repulsion on 3DGS; diversity↑ at similar cost; – kernel/schedule tuning

Abbrev.: 3D Rep. = 3D representation; Guid. = guidance (e.g., SDS, VSD); WGF = Wasserstein gradient flow; 3DGS = 3D Gaussian Splatting.

2.3.1 Probabilistic Reformulations

ProlificDreamer [8] reformulates SDS as *Variational Score Distillation* (VSD), a distributional update under Wasserstein gradient flows:

$$\partial_t \mu_t = \nabla \cdot \big(\mu_t \nabla \tfrac{\delta \mathcal{F}}{\delta \mu}\big). \tag{2.2}$$

This provides principled diversity but at the cost of heavier optimisation and, in many cases, LoRA retraining of the diffusion prior. While distributional approaches are theoretically appealing, their computational burden limits plug-and-play use in modern pipelines. We therefore turn to lighter-weight *kernel-based* mechanisms.

2.3.2 Repulsive Extensions (Kernel-based)

Lightweight approaches introduce diversity by enforcing repulsion among particles. Notably, Repulsive Latent Score Distillation (RLSD) [6] adds a kernelised repulsion term directly in the diffusion latent space, whereas Stein Variational Gradient Descent (SVGD) [18] augments score attraction with an additional kernel field to encourage diversity.

Update-level summary. Both SVGD and RLSD blend score attraction with a kernel-based repulsive field, but SVGD uses a Stein field while RLSD applies a log-sum-exp kernel density term. To make the distinction concrete, their schematic particle updates are:

$$\text{SVGD [18]:} \quad \Delta z_i \;\propto\; \frac{1}{N}\sum_{j=1}^{N}\Big[k(z_j, z_i)\,\nabla_{z_j}\log p(z_j) + \nabla_{z_j} k(z_j, z_i)\Big]. \tag{2.3a}$$

$$\text{RLSD [6]:} \quad \Delta z_i \;\propto\; \nabla_{z_i}\mathcal{L}_{\text{SDS}}(z_i) \;-\; \lambda\,\nabla_{z_i}\log\left(\sum_{j=1}^{N} k(z_i, z_j)\right). \tag{2.3b}$$

From latent to feature space. Both SVGD [18] and RLSD [6] were originally conceived as *latent-space* mechanisms: RLSD introduces kernelised repulsion directly in the diffusion latent z, while SVGD evolves particles under Stein dynamics in latent space. In this work, we lift these mechanisms into *semantic feature space* (e.g., CLIP or DINO embeddings), enabling repulsion to act on semantically meaningful similarities between rendered views rather than raw diffusion latents alone.

Research gap. Existing repulsion methods in text-to-3D have been confined to the *latent space* of SDS (e.g., RLSD with NeRF backbones), while SVGD has not been explored as a diversity mechanism in this domain. Crucially, no prior work has systematically lifted repulsion to *semantic feature embeddings* (e.g., CLIP or DINO) or integrated such mechanisms with the modern *3D Gaussian Splatting (3DGS)* backbone. This leaves open the question of whether kernel-based repulsion in feature space can mitigate collapse while retaining the fidelity and efficiency of SDS-guided 3DGS pipelines.

Table 2.2: Evaluation metrics used in this dissertation, with supporting references.

Metric	References	Rationale / Usage
Fidelity (*CLIP cosine*)	CLIP [23], DreamFusion [7], DreamGaussian [1], ProlificDreamer [8]	Standard image–text alignment score. We report *view-averaged* CLIP similarity across multiple renders to capture text consistency across viewpoints.
Diversity (*feature dissimilarity*)	RLSD [6]	Inter-particle dissimilarity in semantic feature space (CLIP or DINO cosine). Encourages and quantifies set-level diversity beyond fidelity.
Cross-view consistency (*auxiliary*)	Carve3D [17]	Measures similarity of features across different camera views; used as a diagnostic for coherence, not a primary score.
Qualitative evaluation (*human preference*)	DreamGaussian [1], ProlificDreamer [8]	Compact pairwise human studies on fidelity, visual quality, and multi-view consistency provide complementary perceptual validation.

Abbrev.: CLIP = Contrastive Language–Image Pre-training.

2.4 Evaluation in Generative Models

2.4.1 Quantitative Evaluation

Table 2.2 consolidates the evaluation protocol adopted in this dissertation. By aligning with prior work on fidelity, diversity, and diagnostic checks, it ensures that our results are directly comparable while also addressing recent critiques that no single metric suffices for generative evaluation.

No single metric captures fidelity and diversity jointly; recent work therefore advocates multi-metric reporting with sanity checks [24, 25]. In line with this, we report:

Fidelity. Text alignment is measured by CLIP image–text similarity, following DreamFusion [7], DreamGaussian [1], and the original CLIP framework [23]. We report *view-averaged* CLIP fidelity across multiple rendered views, as adopted in recent pipelines [1, 8].

Diversity. We report inter-particle dissimilarity in semantic feature space (CLIP or DINO cosine), following prior repulsion-based methods [6]. For analysis only, we log kernel diagnostics (effective sample size, row-sums) that correlate with collapse.

Cross-view consistency (auxiliary). We additionally compute cross-view similarity scores (DINO-based), as in prior consistency-oriented pipelines [17, 26]. However, since front/back differences are natural in text-to-3D, we treat these metrics as *diagnostic* rather than primary.

Reporting protocol. Metrics are averaged over seeds and view grids, reported as mean±std. Encoders and layers are fixed across methods for fair comparison.

2.4.2 Qualitative Evaluation

Quantitative metrics alone cannot capture all perceptual aspects; in line with prior work (e.g., DreamGaussian and ProlificDreamer), we run a compact *pairwise* human preference study assessing text alignment, visual quality, and multi-view consistency [1, 8]. The protocol, interface, and participant details are provided in Ch. 4.

Summary. Table 2.2 provides an overview of the metrics used in this dissertation, together with their precedent references and intended role. This highlights (i) fidelity via view-averaged CLIP similarity, (ii) diversity via inter-particle dissimilarity in semantic feature space, (iii) cross-view consistency as a supplementary diagnostic, and (iv) human studies as complementary perceptual validation.

2.4.3 Scope of Related Work and Position in the Landscape

Text-to-3D sits alongside several neighbouring lines that target complementary axes to ours. We summarise these briefly to clarify scope and positioning.

High-resolution and mesh-aware pipelines. Beyond DreamFusion-style NeRF optimisation, several systems improve topology or editability by refining to meshes or explicitly encouraging mesh structure (e.g., two-stage optimisation as in Magic3D [9], and mesh-oriented refinements such as Fantasia3D [27]). These advances mainly target output quality and downstream usability rather than inter-sample diversity.

Single-image to 3D priors. Another thread leverages strong 2D priors to lift a single view into 3D (e.g., LRM [10] / One-2-3-45++ [28] / Zero123++ [29]). While they can initialise geometry effectively, their objective differs from text-conditioned multi-sample generation; diversity across seeds is usually not a primary goal.

3DGS-based acceleration and consistency. Replacing NeRF with 3D Gaussian Splatting (3DGS) [2] has enabled efficient differentiable rendering, adopted by DreamGaussian [1] and successors that improve stability and cross-view coherence (e.g., progressive Gaussian optimisation as in GSGEN [30]). These works primarily address efficiency and geometric quality; collapse across seeds can persist because the guidance remains SDS-based.

Consistency, controllability, and relighting. Orthogonal extensions add structure or constraints, such as cross-view consistency modules or relighting control (e.g., Carve3D [17], InstructNeRF2NeRF [31]). They typically improve per-instance fidelity or controllability rather than set-level diversity.

Distributional / probabilistic formulations. ProlificDreamer [8] reframes SDS as VSD under Wasserstein gradient flows, providing a principled route to diversity at the expense of higher compute and, in some cases, LoRA fine-tuning. Such methods indicate that distributional objectives can mitigate collapse, but their overhead limits plug-and-play adoption.

Kernel-based repulsion. Repulsive Latent Score Distillation (RLSD) [6] introduces a lightweight, kernelised repulsion term in DreamFusion-style *latent space*, encouraging separation among particles while retaining SDS attraction. To our knowledge, a systematic study that (i) relocates repulsion to *feature* embeddings and (ii) instantiates it within modern *3DGS* backbones has not been reported. Similarly, although Stein Variational Gradient Descent (SVGD) [18] is well-established in Bayesian inference, its use as a *diversity mechanism* in text-to-3D appears underexplored.

Our scope. Building on these observations, we ask whether repulsion can be lifted from *latent space* into *semantic feature embeddings* (e.g., CLIP/DINO). We study both RLSD and SVGD within a 3DGS backbone, positioning our work as a lightweight, modular extension that targets diversity at comparable fidelity and runtime.

2.5 Summary and Positioning

Together, Tables 2.1 and 2.2 position our contribution within the broader landscape: Table 2.1 situates our method relative to prior pipelines, while Table 2.2 establishes a transparent and multi-faceted evaluation protocol. This dual perspective clarifies both our methodological novelty and the criteria by which it is evaluated.

2.5.1 Key Observations

- With the advent of large text-to-image diffusion models, SDS emerged as a widely adopted approach for text-to-3D, though it suffers from mode collapse and cross-view inconsistency.
- Replacing NeRF with 3DGS improves efficiency and stability, but does not by itself prevent collapse.
- Probabilistic reformulations (e.g., VSD) provide a principled route to diversity, but at substantially higher computational cost.
- Kernel-based repulsion (e.g., RLSD, SVGD) offers lightweight diversity mechanisms, yet prior work applies them only in diffusion *latents* or not at all—leaving open whether feature-space repulsion on 3DGS can mitigate collapse.

Contributions. *This dissertation contributes:*

- **Method novelty:**
  - We instantiate RLSD and SVGD in *semantic feature space* (CLIP/DINO embeddings of renders), rather than diffusion latents.
  - We introduce an N-parallel training scheme in which multiple 3DGS reconstructions are optimised jointly under SDS attraction and feature-space repulsion, enabling diversity to emerge *during* optimisation rather than across independent runs.
- **Controlled comparison:** we present the first head-to-head analysis of SVGD (Stein field) and RLSD (log-sum field) within a common 3DGS baseline with shared renderer, scheduler, and encoders.
- **Empirical validation:** extensive ablations across kernels, feature layers, particle counts, and repulsion schedules demonstrate that feature-space repulsion consistently improves inter-particle diversity without sacrificing fidelity or runtime.

Chapter 3

Methodology

3.1 Concept and Rationale

Goal. For a text prompt y, we optimise N parallel 3D Gaussian Splatting (3DGS) assets ("particles") such that each is *text-aligned* yet *semantically diverse*. Our design preserves the semantic attraction of Score Distillation Sampling (SDS) whilst counteracting mode collapse through a lightweight *feature-space kernel repulsion*.

Design rationale.

- **SDS attraction.** We start from the standard SDS objective for text alignment (Sec. 3.2).
- **Repulsion for diversity.** We add a complementary repulsive field in semantic feature space to discourage particle collapse (Sec. 3.3).
- **Kernels and features.** Repulsion is instantiated with RBF or shifted-cosine kernels on frozen DINOv2 features (Sec. 3.4).
- **Unified objective.** SDS, SVGD, and RLSD fall under one template that cleanly separates *attraction* and *repulsion* (Sec. 3.5).

Rationale. The method is deliberately designed around lightweight, feature-space repulsion. Semantic embeddings such as DINOv2 provide representations that capture object- and part-level variation, which are essential for both quantifying and promoting diversity. Kernel-based repulsion offers a tractable and computationally efficient mechanism that can be incorporated into existing SDS pipelines without modification to the diffusion backbone. Moreover, the N-parallel formulation ensures that diversity arises jointly during optimisation rather than as a by-product of stochastic variation. Overall, the design aims to enhance semantic diversity whilst maintaining fidelity and computational efficiency.

The overall system pipeline is illustrated in Figure 3.1, which highlights the separation between *attraction* via SDS (Sec. 3.2) and *repulsion* in feature space (Secs. 3.3 and 3.4). This figure serves as a reference for the unified objective introduced later (Sec. 3.5; see also Equations (3.7) to (3.9)).

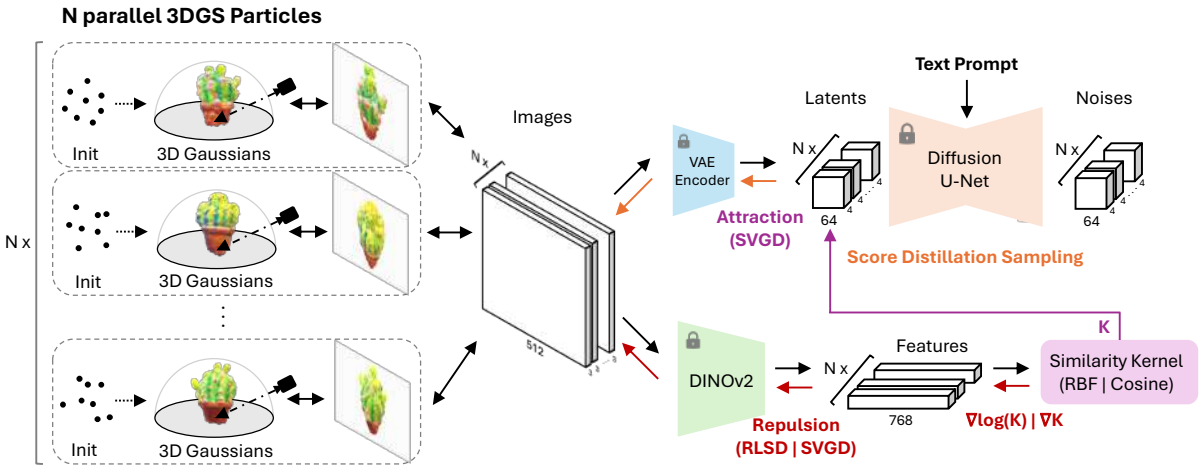


Figure 3.1: **System overview.** For each prompt we optimise N 3DGS particles in parallel [2]. *Attraction* (top) follows standard SDS: render $I^{(i)} \to z_0^{(i)} \to z_t^{(i)}$ and obtain the residual $g^{(i)}$ from the frozen Stable Diffusion U-Net (Sec. 3.2). *Repulsion* (bottom) operates in semantic feature space: extract $\mathbf{f}^{(i)} = \phi_\ell(I^{(i)})$, construct a positive kernel K (RBF or shifted-cosine), and compute a repulsive field $\mathbf{R}_i$ as either SVGD (unnormalised Stein) or RLSD (log-sum normalised) (Secs. 3.3 and 3.4). The *unified objective* combines attraction and repulsion with weights λ_{sd} and λ_{rep} (Equations Equation (3.7)–Equation (3.9)); updates are restricted to the 3DGS parameters, while diffusion and feature encoders remain frozen. An optional "same-noise" mode isolates the causal effect of repulsion by sharing diffusion noise across particles.

3.2 SDS Attraction (baseline)

Let $I = \mathcal{R}(\mathcal{G}, \mathbf{c})$ be a render of a 3DGS scene $\mathcal{G}$ at camera pose $\mathbf{c}$. The Stable Diffusion VAE encoder gives $z_0 = \mathcal{E}(I)$ [19]. For diffusion timestep t and noise ϵ,

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \hat{\epsilon} = \epsilon_\theta(z_t, t, y). \tag{3.1}$$

The SDS residual and scalar energy are

$$g \;=\; w(t)\,(\hat{\epsilon} - \epsilon), \quad w(t) = 1 - \bar{\alpha}_t, \tag{3.2}$$

$$\mathcal{L}_{\text{att}} = \tfrac{1}{2}\|g\|_2^2, \tag{3.3}$$

implemented as an MSE to the stop-gradient target $(z_0 - g)$.

3.3 Repulsive Fields for Diversity

We seek a repulsive field $\mathbf{R}_i$ in semantic feature space to discourage coalescence. Let $\mathbf{f}^{(i)}$ denote the feature of particle i and let $K = [K_{ij}]$ be a positive kernel over features. We consider two repulsive fields: SVGD [18] and RLSD [6]. We consider two alternatives for the repulsive field—SVGD and RLSD—formalised below:

$$\mathbf{R}_i^{\text{SVGD}} = \nabla_{\mathbf{f}^{(i)}} \sum_j K_{ij}, \qquad \mathbf{R}_i^{\text{RLSD}} = \nabla_{\mathbf{f}^{(i)}} \log\Big(\sum_j K_{ij}\Big). \tag{3.4}$$

SVGD gives an unnormalised Stein field (transport view); RLSD gives a mass-normalised log-sum field (potential view). For an RBF kernel,

$$\mathbf{R}_i^{\text{SVGD}} = \tfrac{2}{h}\sum\nolimits_j K_{ij}(\mathbf{f}^{(j)} - \mathbf{f}^{(i)}), \quad \mathbf{R}_i^{\text{RLSD}} = \mathbf{R}_i^{\text{SVGD}} / (\sum_j K_{ij} + \varepsilon).$$

In Equation (3.4), $\mathbf{R}^{\text{SVGD}}$ is an unnormalised Stein field (transport view), whereas $\mathbf{R}^{\text{RLSD}}$ applies a log-sum normalisation (potential view). This distinction later motivates our choice of RLSD with an RBF kernel as the default configuration.

3.4 Semantic Features and Kernels

We extract frozen DINOv2 features at layer ℓ [5]: $\mathbf{f}^{(i)} = \phi_\ell(I^{(i)}) \in \mathbb{R}^d$.

$$\text{RBF:} \quad K_{ij} = \exp\Big(-\frac{\|\mathbf{f}^{(i)}-\mathbf{f}^{(j)}\|_2^2}{h}\Big), \quad h = \frac{\text{median}_{p\neq q}\|\mathbf{f}^{(p)}-\mathbf{f}^{(q)}\|_2^2}{\log N} \cdot \frac{1}{\beta_{\text{rbf}}}, \tag{3.5}$$

$$\text{Shifted-cosine:} \quad K_{ij} = \Big(\tfrac{1-\varepsilon_s}{2}\big(\cos(\mathbf{f}^{(i)},\mathbf{f}^{(j)})+1\big)+\varepsilon_s\Big)^{\beta_{\cos}}, \tag{3.6}$$

where the small shift ε_s ensures strict positivity (required by RLSD). The temperatures β_{rbf} and $\beta_{\cos}$ control kernel sharpness in their respective formulations. All cosine similarities are computed after ℓ_2-normalising feature vectors. Here, the median heuristic sets the base scale, and the division by $\log N$ tempers bandwidth growth with the number of particles in high-dimensional settings, preventing overly flat kernels as N increases.

We now unify attraction and repulsion into a single objective, exposing the minimal changes required to turn SDS into SVGD or RLSD; see Equations (3.7) to (3.9).

3.5 Unified Objective (attraction $\oplus$ repulsion)

For $m \in \{\text{SDS, SVGD, RLSD}\}$ the total objective decomposes as

$$\mathcal{L}^{(m)} \;=\; \lambda_{\text{sd}}\,\mathcal{L}^{(m)}_{\text{attr}} \;+\; \lambda_{\text{rep}}\,\mathcal{L}^{(m)}_{\text{rep}}. \tag{3.7}$$

Attraction (common form with method-specific guidance). Let $g^{(i)}$ be the SDS residual and K_{ij} the (detached) kernel entries:

$$\mathcal{L}^{(m)}_{\text{attr}} \;=\; \frac{1}{2N}\sum_{i=1}^{N}\Big\|\underbrace{\begin{cases} g^{(i)}, & m=\text{SDS},\\ \sum_{j=1}^{N}\text{sg}[K_{ij}]\,g^{(j)}, & m=\text{SVGD},\\ g^{(i)}, & m=\text{RLSD}\end{cases}}_{\text{attraction guidance for particle } i}\Big\|_2^2. \tag{3.8}$$

Repulsion (field–feature coupling). With DINO features $\mathbf{f}^{(i)}$ and fields from Eq. (3.4),

$$\mathcal{L}^{(m)}_{\text{rep}} \;=\; \frac{1}{N}\sum_{i=1}^{N}\Big\langle\underbrace{\begin{cases}\mathbf{0}, & m=\text{SDS},\\ \text{sg}[\mathbf{R}_i^{\text{SVGD}}], & m=\text{SVGD},,\\ \text{sg}[\mathbf{R}_i^{\text{RLSD}}], & m=\text{RLSD}\end{cases}}_{\text{repulsive field (stop-grad)}},\ \mathbf{f}^{(i)}\Big\rangle. \tag{3.9}$$

This stop-gradient treatment ensures that repulsive fields are treated as fixed reference directions, so that gradients flow only through the feature embeddings. In practice, this prevents degenerate feedback loops where the field itself would adapt to cancel the repulsion signal.

A full training step is summarised in Alg. 1, integrating SDS attraction with feature-space repulsion. Only 3DGS parameters are updated; the diffusion model, VAE, and feature extractor remain frozen throughout training.

3.6 Algorithm

Algorithm 1 One training step with SDS attraction and feature-space repulsion. Gradients update 3DGS parameters, while diffusion, VAE, and feature encoders remain frozen.

Input: Particles $\{\mathcal{G}^{(i)}\}_{i=1}^{N}$, prompt y, iteration s

Sample orbit camera $\mathbf{c}$ and scheduled resolution (H, W)
for $i = 1 \dots N$ **do**
  $I^{(i)} \leftarrow \mathcal{R}(\mathcal{G}^{(i)}, \mathbf{c})$ // 3DGS render
  $\mathbf{f}^{(i)} \leftarrow \phi_\ell(I^{(i)})$ // DINOv2 feature
  $z_0^{(i)} \leftarrow \mathcal{E}(I^{(i)})$ // SD VAE latent
  Sample $t \sim \mathcal{U}[t_{\min}, t_{\max}]$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
  $z_t^{(i)} \leftarrow \sqrt{\bar{\alpha}_t}\, z_0^{(i)} + \sqrt{1-\bar{\alpha}_t}\, \epsilon$
  $\hat{\epsilon}^{(i)} \leftarrow \epsilon_\theta(z_t^{(i)}, t, y)$
  $g^{(i)} \leftarrow w(t)\left(\hat{\epsilon}^{(i)} - \epsilon\right)$ // cf. Eq. (3.2)
end for
Build kernel $K \in \mathbb{R}^{N\times N}$ on $\{\mathbf{f}^{(i)}\}$ (RBF or shifted-cosine)
Compute repulsion fields $\mathbf{R}_i$ via Eq. (3.4)
Compute $\mathcal{L}_{\text{attr}}^{(m)}$ by Eq. (3.8) and $\mathcal{L}_{\text{rep}}^{(m)}$ by Eq. (3.9)
Total loss $\mathcal{L}^{(m)} = \lambda_{\text{sd}}\, \mathcal{L}_{\text{attr}}^{(m)} + \lambda_{\text{rep}}\, \mathcal{L}_{\text{rep}}^{(m)}$; update 3DGS parameters only
Periodically densify/prune Gaussians; optionally reset opacities

A full training step is summarised in Algorithm 1, integrating SDS attraction with feature-space repulsion. During optimisation, the diffusion backbone and feature extractor remain fixed, so gradients update only the 3DGS parameters.

With the objective and optimisation loop in place, we next motivate the design choices that make the method lightweight and practical (Section 3.7).

3.7 Design Decisions and Alternatives

Our methodology was shaped by several critical design choices. In each case, we considered alternatives documented in the background (Ch. 2) and adopted the option that best balanced fidelity, diversity, and computational efficiency.

3.7.1 Choice of 3D Representation

We adopted **3D Gaussian Splatting (3DGS)** as the rendering backbone instead of NeRF-based methods. While NeRFs [13, 7] offer photorealistic rendering, their slow optimisation and high compute cost make them impractical for multi-particle training. 3DGS [2, 1] provides real-time rasterisation and stable gradients, allowing us to scale to N-parallel optimisation runs.

3.7.2 Location of Repulsion

Repulsive fields could in principle act in diffusion *latents*, semantic *features*, or distributional spaces. We focused on **feature-space repulsion** (CLIP/DINO embeddings) because: (i) latent-space repulsion (e.g. RLSD [6]) risks operating on representations that are less semantically aligned, (ii) distributional methods (e.g. VSD [8]) provide principled guarantees but incur significantly higher compute, and (iii) feature embeddings directly capture semantic similarity between renders.

3.7.3 Kernel and Field Variants

We evaluated both **RBF kernels** and **shifted-cosine kernels**. RBF kernels provide smooth isotropic repulsion but require careful bandwidth tuning; cosine kernels better align with angular similarity in high-dimensional features. For the repulsive field, we compared **SVGD** (unnormalised Stein dynamics)

and **RLSD** (log-sum normalised), as detailed in Sec. 3.3. This dual design allowed us to disentangle kernel effects from field normalisation.

3.7.4 Parallel vs. Independent Training

Rather than training N independent models and comparing outputs post hoc, we introduced an **N-parallel optimisation scheme**. This design allows repulsion to shape the trajectories of particles *during* training, ensuring diversity emerges jointly rather than being an incidental by-product. Although parallel training increases per-step memory usage, it amortises compute and provides stronger guarantees against collapse.

3.7.5 Noise Sharing for Causal Isolation

To isolate the contribution of repulsion, we introduced an option to **share diffusion noise across particles**. This controlled setting (`force_same_noise=True`) ensures that observed diversity cannot be attributed to stochastic variation alone. In practice, we evaluate both settings: the shared-noise variant is used in early ablations to reveal the causal effect of repulsion ($\text{Exp0}_{\text{shared}}$), while the independent-noise variant ($\text{Exp0}_{\text{indep}}$) reflects realistic usage and serves as the default in subsequent comparisons (see Section 4.3 and Chapter 4).

Summary. These design choices—3DGS backbone, feature-space repulsion, kernel/field variants, parallel optimisation, and controlled stochasticity— jointly define the methodological scope. Each decision reflects a trade-off between alternative approaches discussed in Ch. 2, prioritising computational efficiency and robustness of semantic diversity without compromising fidelity. In the next chapter (Ch. 4), we validate these design choices through systematic experiments, first isolating the causal effects of repulsion (with and without noise sharing), and then evaluating performance across kernels, fields, and training setups. This progression from methodological rationale to empirical evidence ensures that our contribution is both principled and experimentally grounded.

Guiding principle. The methodology reflects a principle of minimal yet effective modification: only the components strictly required to mitigate mode collapse are introduced, whilst preserving efficiency, modularity, and reproducibility. This ensures that the proposed mechanism remains extensible and broadly applicable across future text-to-3D frameworks, consistent with the dissertation's emphasis on reproducible and generalisable research practice.

Chapter 4

Evaluation

4.1 Overview

Our evaluation targets *semantic diversity* in text-to-3D generation, while monitoring *text fidelity* and *cross-view consistency* as diagnostics. The central question is whether introducing a feature-space repulsion can increase semantic diversity across 3D samples without sacrificing fidelity or geometric coherence.

To address this, we proceed in four stages: (i) metrics definition; (ii) experimental setup (training, rendering, prompts, seeds); (iii) ablation studies across repulsion mechanism, strength, guidance, and kernel temperature; (iv) final comparison and extended analysis (human evaluation, efficiency, limitations).

In particular, Experiments 1–5 vary individual hyperparameters to identify favourable settings, and Experiment 6 consolidates the best configuration and benchmarks it against the matched multi-particle baseline (Sec. 4.3).

This chapter follows this roadmap: we begin with controlled ablations (Exps1–5), synthesise their findings into a consolidated configuration, and then conduct the final comparison (Exp6) followed by extended analyses.

Rationale and principles. The evaluation is designed to isolate the causal effect of feature-space repulsion under controlled conditions. First, we adopt within-pipeline, like-for-like ablations to avoid confounding architectural factors; all runs share training schedules, seeds, and camera protocols. Second, we report multi-view metrics with explicit aggregation across particles and viewpoints, coupled with a consistency guard to discourage pathological solutions. Third, model selection follows a diversity-weighted rule with an ε-consistency constraint, prioritising semantic spread whilst preserving geometric coherence. Finally, we complement automatic metrics with a small-scale human study and a representation-space analysis (PCA), providing convergent evidence consistent with reproducible and generalisable research practice.

4.2 Experimental Setup

4.2.1 Prompts

The evaluation prompts were selected to ensure both comparability with prior work and coverage across representative evaluation dimensions. Four prompts (`hamburger`, `cactus`, `tulip`, `icecream`) are drawn from DreamGaussian [1], which primarily emphasised efficiency and fidelity. Two further prompts (`bulldozer`, `sundae`) were adopted from RLSD [6], a method designed to highlight diversity. This selection therefore enables our study to inherit established baselines in fidelity and efficiency, while also incorporating diversity-oriented test cases.

In addition, the chosen prompts span complementary axes: artificial versus natural objects, structural versus appearance-focused variation, and canonical versus extended benchmarks. For brevity, we refer to prompts by their `key` (e.g. `bulldozer`, `cactus`) throughout the remainder of the dissertation; the exact key–value mappings are provided in Table 4.1.

Table 4.1: Prompts used across experiments, with provenance, exact text, and evaluation rationale. DreamGaussian [1] focused on efficiency/fidelity, while RLSD [6] targeted diversity.

Key	Source	Exact text	Rationale
`bulldozer`	RLSD	`"a bulldozer made out of toy bricks"`	Artificial; strong structural composition (toy bricks); viewpoint-sensitive.
`cactus`	DreamGaussian	`"a small saguaro cactus planted in a clay pot"`	Natural; object–context composition (plant + pot); structural consistency.
`tulip`	DreamGaussian	`"a photo of a tulip"`	Natural; fine-grained texture (petals); used across fidelity/diversity pipelines.
`icecream`	DreamGaussian	`"a photo of an ice cream"`	Artificial; high appearance variability (colour, toppings); viewpoint-invariant.
`hamburger`	DreamGaussian	`"a photo of a hamburger"`	Canonical SDS benchmark; illustrates mode collapse.
`sundae`	RLSD	`"an ice cream sundae"`	Artificial; variant of `icecream`; tests compositional generalisation.

4.2.2 Training

We train for 1k iterations with progressive render resolutions (128→256→512). Guidance = 50 unless varied (Exp4). Repulsion strength λ=1000 unless swept (Exp2–3). All runs freeze UNet/CLIP weights. Comprehensive defaults—including densification, opacity, and logging settings—are listed in Sec. B.1.

Score-distillation call. We use a DDIM schedule as in DreamGaussian. To ensure a fair comparison, both the baseline and our method use the same noise protocol: timestep t and diffusion noise ϵ are synchronised across all particles (`force_same_t`, `force_same_noise`). This guarantees that any difference in diversity arises solely from the presence or absence of the repulsion term. For completeness, we also report results under an independent-noise setting in the appendix. UNet/CLIP/Tokenizer weights remain frozen throughout.

Feature space. We extract embeddings from the [CLS] token of DINOv2. Because our objective is to promote semantic diversity, the main experiments use `feature_layer='last'`, which captures the most abstract semantic features. Additional ablations across `early/mid/last` layers are reported in the appendix.

Logging. Quantitative metrics (fidelity, diversity, consistency) are recorded every 50 iterations, while losses and efficiency statistics (wall/GPU time, sub-component timings, memory usage, throughput) are logged every 10 iterations. Unless otherwise specified, analyses are reported with respect to the final iteration. The full logging configuration is detailed in Sec. B.1.

Configuration and reproducibility. To facilitate systematic ablations and ensure reproducibility, all experiments are driven by modular `.yaml` configuration files. Key hyperparameters (e.g., repulsion strength, guidance scale, kernel temperature), renderer options, and logging schedules can be altered without code changes, enabling transparent reporting and efficient exploration of the design space. For each run, the exact configuration is version-controlled and archived alongside checkpoints and metrics logs, with a configuration hash recorded in the run metadata. This configuration-first protocol preserves the full experimental context and materially improves comparability and repeatability across studies.

4.2.3 Rendering

Evaluation is conducted using V=8 fixed viewpoints (zero elevation; azimuths uniformly distributed over $[0, 360)$). The camera radius is prompt-specific to prevent cropping (e.g. `bulldozer`: 4.5; `hamburger`: 3.0; `icecream/cactus/tulip/sundae`: 4.0). Training views are randomly sampled, whereas evaluation views are fixed and shared across methods to enable paired comparisons.

Multi-view evaluation. A dedicated `Visualizer` module was implemented to render the fixed V=8 azimuth views. The resulting multi-view images are used both to compute CLIP/DINO embeddings for evaluation metrics and to obtain DINOv2 embeddings for repulsion. This design decouples metric computation from training-time sampling, ensuring that all reported scores are derived under identical camera settings and view distributions.

Extended visualisation. For the final best-performing model (`exp6_ours_best`), we additionally rendered V=120 azimuth views to produce 360° turntable animations (saved as GIFs), following the standard practice in text-to-3D works [7, 9, 8]. The optimised Gaussians were also exported as Blender-compatible meshes, enabling interactive inspection of the generated assets.

4.2.4 Metrics: Fidelity, Diversity, Consistency

Our evaluation targets *semantic diversity* as the primary objective, while preserving *text fidelity*; *cross-view consistency* serves as a diagnostic. All embeddings are ℓ_2-normalised before computing similarities. For evaluation, we use DINO features [32] for diversity and consistency, and CLIP [23] (OpenCLIP implementation [33]) for text–image fidelity. By contrast, the proposed semantic repulsion is implemented in a DINOv2 feature space [5], whose improved semantic encoding better supports our diversity objective.

Fidelity ($\mathcal{F}\uparrow$): multi-view CLIP

We measure text–image fidelity with OpenCLIP ViT-bigG-14 [23, 33]. Scores are cosine similarities between rendered images and the prompt embedding, aggregated over particles and views:

$$\mathcal{F} = \frac{1}{N}\sum_{p=1}^{N}\frac{1}{V}\sum_{v=1}^{V}\left\langle \phi_{\text{clip}}(I_v^{(p)}),\, \psi_{\text{clip}}(t)\right\rangle. \tag{4.1}$$

Diversity ($\mathcal{D}\uparrow$): inter-particle spread

Diversity follows the RLSD literature [6], using DINO embeddings [32]. We compute one minus the mean pairwise cosine per view, then average across views:

$$\mathcal{D} = \frac{1}{V}\sum_{v=1}^{V}\left(1 - \frac{2}{N(N-1)}\sum_{i<j}\left\langle \phi_{\text{dino}}(I_v^{(i)}),\, \phi_{\text{dino}}(I_v^{(j)})\right\rangle\right). \tag{4.2}$$

Consistency ($\mathcal{C}\uparrow$): within-particle agreement

Consistency captures whether each particle retains appearance across views. It is defined as the mean cosine similarity across all view pairs of the same particle:

$$\mathcal{C} = \frac{1}{N}\sum_{p=1}^{N}\frac{1}{V(V-1)}\sum_{v_1\neq v_2}\left\langle \phi_{\text{dino}}(I_{v_1}^{(p)}),\, \phi_{\text{dino}}(I_{v_2}^{(p)})\right\rangle. \tag{4.3}$$

Variants. We additionally considered LPIPS [34], but it introduces substantial VRAM overhead (due to extra forward passes through a perceptual network) without improving semantic sensitivity, so we omit it from the main analysis. We also briefly experimented with HuggingFace CLIP instead of OpenCLIP, since the official DreamGaussian repository adopts HF CLIP by default, but found no consistent benefit.

Statistical reporting and conventions. Unless otherwise specified, we report means ± standard error (SE) aggregated over prompts × seeds. Significance is assessed using paired t-tests for within-condition comparisons—across seeds for model settings and across participants for human-study ratings—with Bonferroni correction where appropriate [35, 36]. We also present Pareto views when interpreting fidelity–diversity (and related) trade-offs.

4.2.5 Seeds for Reproducibility

Unless otherwise noted, results are averaged over seeds $\{42, 123, 456, 789\}$; Exp2 uses seed 42 only. Particle j is initialised with $(\texttt{seed} + 10^6 \times j)$, and score-distillation noise uses an independent `torch.Generator`.

Figure 4.1: **Baseline (wo)**: N=8 particles without repulsion, trained with shared timestep and noise per SDS call (`force_same_t=True`, `force_same_noise=True`). Example fixed-view collage (2×4, 0° az., 0° el.) for the `cactus` prompt (seed 42). This run is representative of the multi-particle baseline used throughout.

4.3 Baselines

Definition and settings. Unless otherwise stated, all ablations are compared against a matched *multi-particle baseline* with N=8 parallel 3DGS particles and no repulsion term (`repulsion_type=none`). We synchronise the timestep and diffusion noise across particles (`force_same_t=True`, `force_same_noise=True`) to decouple repulsion from stochasticity. Training schedule and renderer follow Sec. 4.2, while default hyperparameters are listed in Sec. B.1. Seed protocol follows Sec. 4.2.5.

Rationale. This baseline preserves the standard SDS attraction while disabling feature-space repulsion, thereby serving as a direct control to isolate the effect of the proposed mechanism. With shared noise, the baseline typically collapses to near-identical shapes, while repulsive variants successfully separate trajectories. Giving particles independent (t, ϵ) increases baseline diversity modestly, but still far below our method (Table B.3).

Scope. We deliberately restrict comparisons to within-pipeline ablations, avoiding cross-backbone baselines such as DreamFusion or ProlificDreamer, which would conflate architectural differences with the repulsion mechanism. Our evaluation therefore directly addresses the central hypothesis: feature-space repulsion can enhance semantic diversity while preserving fidelity and cross-view consistency, thereby mitigating the mode collapse observed in SDS-based text-to-3D generation [7, 6].

Having established this baseline, we now turn to controlled ablations. In Experiments 1–5, we vary one hyperparameter axis at a time to investigate the role of repulsion (mechanism, strength, guidance, and kernel temperature), before consolidating the best-performing configuration and benchmarking it against the baseline in Experiment 6.

4.4 Experimental Axes

- **Exp1** (Repulsion & kernel): `repulsion_type` $\in$ {rlsd, svgd}, `kernel_type` $\in$ {cosine, rbf}.
- **Exp2** (Coarse λ): $\lambda \in \{1, 10, 100, 1000, 10000\}$ (seed 42).
- **Exp3** (Fine λ): $\lambda \in \{600, 800, 1000, 1200, 1400\}$.

- **Exp4** (Guidance): `guidance_scale` $\in \{30, 50, 70, 100\}$.
- **Exp5** (RBF β): $\beta \in \{0.5, 1.0, 1.5, 2.0\}$.
- **Final (Exp6)**: best from Exp1–5 vs. matched baseline.

Across ablations, each experiment varies a single axis while holding others fixed. Exp6 consolidates the best-performing settings from Exp1–5 (using the rule in Sec. 4.4.1) and compares them against the matched baseline. The objective is to identify configurations that maximise semantic diversity $\mathcal{D}$ while maintaining text fidelity $\mathcal{F}$ and avoiding excessive degradation in cross-view consistency $\mathcal{C}$.

4.4.1 Model Selection: Diversity-weighted Rule

To operationalise this selection, we adopt a simple Pareto-inspired scalarisation rather than a full evolutionary search. Each candidate $c \in \mathcal{C}$ has scores $(\mathcal{F}_c, \mathcal{D}_c, \mathcal{C}_c)$, which we min–max normalise $(\mathcal{F}, \mathcal{D})$ within the sweep to remove scale effects:[1]

$$\tilde{\mathcal{F}}_c = \frac{\mathcal{F}_c - \min \mathcal{F}}{\max \mathcal{F} - \min \mathcal{F} + \delta}, \quad \tilde{\mathcal{D}}_c = \frac{\mathcal{D}_c - \min \mathcal{D}}{\max \mathcal{D} - \min \mathcal{D} + \delta}.$$

We then rank by the weighted Euclidean distance to the utopia point $(1, 1)$,

$$U(c) \;=\; \sqrt{w_{\text{fid}}\,(1 - \tilde{\mathcal{F}}_c)^2 \;+\; w_{\text{div}}\,(1 - \tilde{\mathcal{D}}_c)^2}\,,$$

using $(w_{\text{fid}}, w_{\text{div}}) = (0.40, 0.60)$ to privilege diversity. To exclude pathological solutions, let $\mathcal{C}_{\max} = \max_{c \in \mathcal{C}} \mathcal{C}_c$ and retain only c with $\mathcal{C}_c \geq \mathcal{C}_{\max} - \varepsilon$ (default $\varepsilon = 0.02$). The selected configuration is

$$c^\star \;=\; \arg \min_{c \in \mathcal{C}:\, \mathcal{C}_c \geq \mathcal{C}_{\max} - \varepsilon} U(c).$$

This weighting scheme privileges diversity while enforcing near-maximum consistency, reflecting our emphasis on mitigating collapse without sacrificing structural coherence.

This follows standard multi-objective practice [37, 38]. More elaborate Pareto-elitist searches (e.g., NSGA-II [39]) could be applied, but our lightweight scalarisation suffices given the low dimensionality of the sweep space (at most 2–5 hyperparameters at a time).

Algorithm 2 Diversity-weighted selector with ε-consistency

Input: Candidate set $\mathcal{C}$ with $(\mathcal{F}_c, \mathcal{D}_c, \mathcal{C}_c)$; weights $w_{\text{fid}}, w_{\text{div}}$; tolerance ε

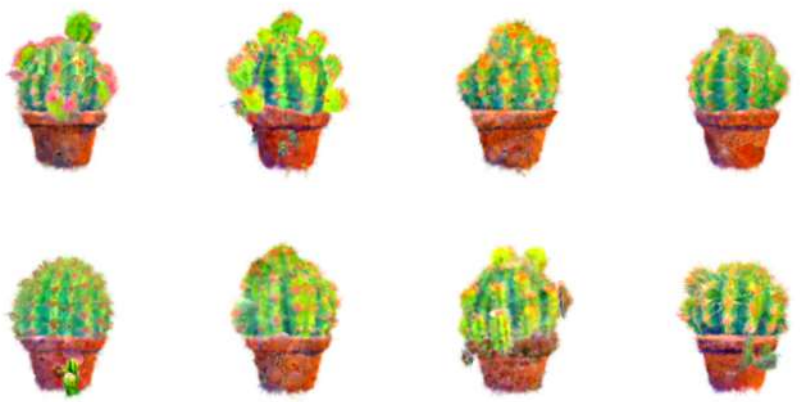
1: Normalise $\tilde{\mathcal{F}}_c, \tilde{\mathcal{D}}_c$ over $\mathcal{C}$ (min–max with δ)
2: $\mathcal{C}_{\max} \leftarrow \max_{c \in \mathcal{C}} \mathcal{C}_c$; $\quad \mathcal{S} \leftarrow \{c \in \mathcal{C} : \mathcal{C}_c \geq \mathcal{C}_{\max} - \varepsilon\}$
3: For $c \in \mathcal{S}$, compute $U(c) = \sqrt{w_{\text{fid}}(1 - \tilde{\mathcal{F}}_c)^2 + w_{\text{div}}(1 - \tilde{\mathcal{D}}_c)^2}$
4: Break ties by higher $\mathcal{C}_c$; if still tied, higher $\mathcal{F}_c$
5: **return** $c^\star = \arg\min_{c \in \mathcal{S}} U(c)$

The complete selection procedure is summarised in Algorithm 2, which formalises our diversity-weighted rule and ensures that the chosen configuration balances fidelity and diversity while maintaining near-maximum consistency.

Building on the experimental setup (Sec. 4.2), rendering protocol (Sec. 4.2.3), and evaluation metrics (Sec. 4.2.4), we now conduct controlled ablations across the experimental axes (Exps 1–5). Each ablation varies a single factor while applying the diversity-weighted selection rule (Sec. 4.4.1), and results are compared against the matched multi-particle baseline (Sec. 4.3). We begin with the repulsion mechanism and kernel in Exp1.

[1] $\delta = 10^{-8}$ is added only to avoid division-by-zero in degenerate sweeps.

(a) **SVGD–COS**: narrowest dispersion; limited morphological diversity.



(b) **SVGD–RBF**: broader than COS, yet less diverse than RLSD.

(c) **RLSD–COS**: improved dispersion with strong prompt fidelity.

(d) **RLSD–RBF**: richest semantic diversity (pot/branching/blossoms) with strong alignment.

Figure 4.2: **Exp1 — Fixed-view qualitative comparison across repulsion mechanism × kernel (RLSD/SVGD × COS/RBF; *cactus*, seed 42; az. 0°, el. 0°).** Each panel shows a 2×4 grid from the same camera. RLSD–RBF exhibits the broadest morphological and compositional diversity, whereas SVGD–COS shows the narrowest dispersion.

Table 4.2: **Exp1 (repulsion mechanism × kernel; with baseline).** Means ± SE over prompts × seeds (N=8, V=8). The multi-particle baseline without repulsion (wo) is included for comparison. RLSD improves diversity over SVGD at near-constant fidelity, and RBF kernels dominate cosine for both mechanisms.

Method	$\mathcal{F}\uparrow$	$\Delta\mathcal{F}$	$\mathcal{D}\uparrow$	$\Delta\mathcal{D}$	$\mathcal{C}\uparrow$	$\Delta\mathcal{C}$
Baseline (wo)	0.398 ± 0.001	–	0.138 ± 0.002	–	0.853 ± 0.001	–
SVGD–COS	0.369 ± 0.001	-0.030	0.181 ± 0.002	+0.043	0.846 ± 0.002	-0.007
SVGD–RBF	0.380 ± 0.000	-0.018	0.254 ± 0.002	+0.116	0.835 ± 0.001	-0.019
RLSD–COS	0.393 ± 0.000	-0.005	0.198 ± 0.002	+0.060	0.845 ± 0.002	-0.008
RLSD–RBF	0.388 ± 0.000	-0.011	0.267 ± 0.002	+0.129	0.831 ± 0.001	-0.022

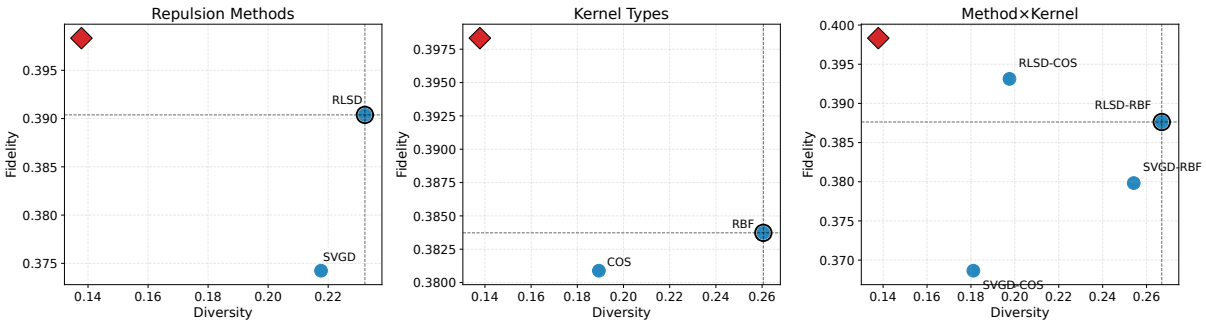


Figure 4.3: **Exp1 — Decomposed fidelity–diversity comparisons (means over prompts × seeds).** *Left:* repulsion mechanism (SVGD → RLSD) at a fixed kernel (RBF). *Centre:* kernel type (COS → RBF) at a fixed mechanism (RLSD). *Right:* full method × kernel grid. RLSD shifts points rightwards (greater diversity) at near-constant fidelity, and RBF dominates COS. The red diamond marks the multi-particle baseline (wo).

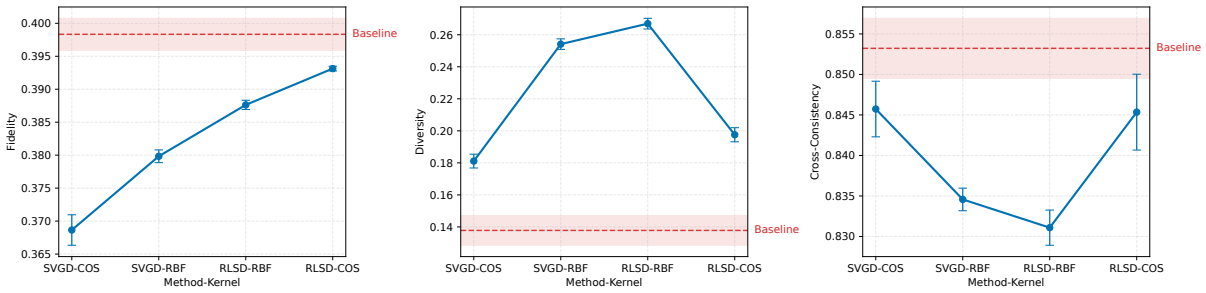


Figure 4.4: **Exp1 — Repulsion–kernel ablation (means ± SE over prompts × seeds).** Three panels report fidelity (left), semantic diversity (centre; primary), and cross-view consistency (right; diagnostic). Red dashed bands indicate the multi-particle baseline (wo). RLSD–RBF yields the strongest diversity gains while keeping fidelity close to baseline; consistency remains within a narrow band across variants.

4.4.2 Exp1: Repulsion Mechanism and Kernel

Objective. We study how the *repulsion mechanism* (SVGD vs. RLSD) and the *kernel* (Shifted-cosine (COS) vs. RBF) affect the trade-off between semantic diversity (primary) and text fidelity, while also monitoring cross-view consistency.

Protocol. We evaluate four configurations—SVGD–COS, SVGD–RBF, RLSD–COS, RLSD–RBF—under identical settings (λ=1000, β=1.0, guidance = 50, N=8, V=8); see Sec. 4.2. Metrics are averaged over prompts×seeds $\{42, 123, 456, 789\}$. The multi-particle baseline (wo) is included for comparison (Table 4.2).

Results. Quantitatively (Table 4.2), **RLSD** improves diversity over SVGD at near-constant fidelity, and **RBF** dominates COS for both mechanisms. Relative to the baseline ($\mathcal{F}$=0.398, $\mathcal{D}$=0.138), **RLSD–RBF**

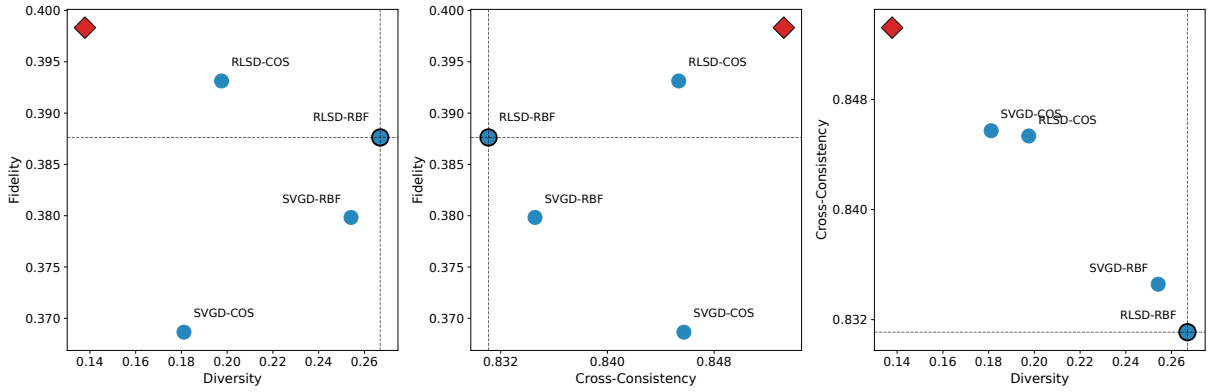


Figure 4.5: **Exp1 — Pareto views of the trade-offs (means over prompts × seeds).** The desirable region is higher diversity at equal (or higher) fidelity. **RLSD–RBF** lies on the Pareto frontier—offering strictly higher diversity at matched fidelity—whereas `RLSD-COS` and `SVGD-COS` are dominated. The red diamond denotes the multi-particle baseline (wo).

reaches $\mathcal{D}$=0.267 (Δ+0.129; $\sim$93% $\uparrow$) with a small fidelity change ($\Delta\mathcal{F} = -0.011$). The decomposed views (Figure 4.3) show: switching SVGD $\rightarrow$ RLSD at fixed kernel shifts points rightward (higher diversity) with minimal fidelity change (left); switching COS $\rightarrow$ RBF at fixed mechanism yields another rightward gain (centre). The full grid (right) and the Pareto plots (Figure 4.5) place **RLSD–RBF** on the frontier. Fixed-view grids in Figure 4.2 visually corroborate these trends.

Interpretation. Mechanism and kernel provide *complementary, roughly additive* diversity gains: RLSD supplies a mechanism-level push, while RBF supplies a kernel-level push. Fidelity deltas are small across RLSD variants, indicating the gains target semantic spread rather than sacrificing prompt adherence. COS under both mechanisms shows the narrowest spread (also visible in Figure 4.2).

Conclusion. Among all combinations, **RLSD–RBF** best satisfies our objective—maximising semantic diversity at near-baseline fidelity with stable cross-view consistency. We adopt RLSD–RBF as the default repulsion configuration for subsequent ablations.

4.4.3 Exp2: Coarse Sweep of Repulsion Strength λ

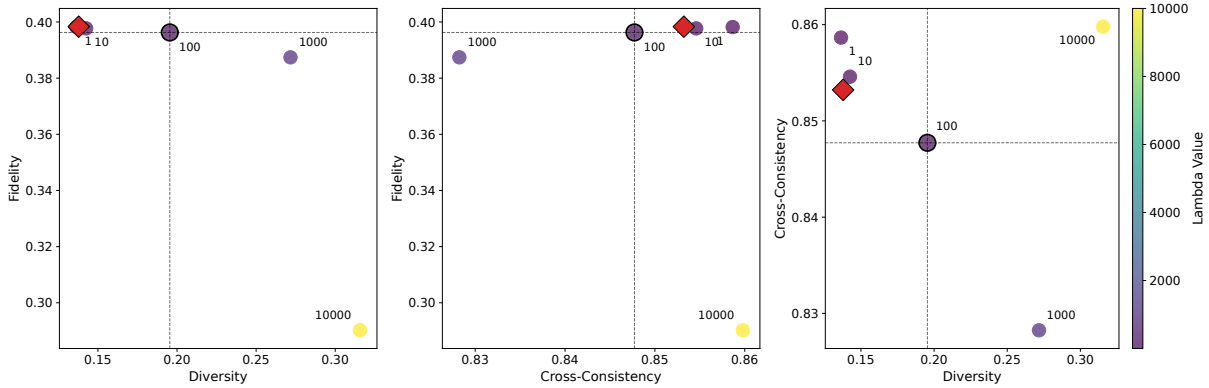
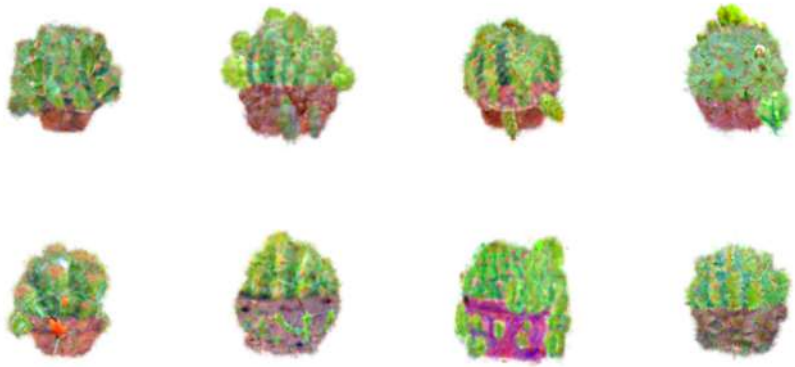


Figure 4.6: **Exp2 — Coarse sweep of repulsion strength λ (Pareto view; RLSD–RBF; *cactus*; seed 42).** Points are means over prompts (single seed). Diversity rises monotonically with λ while fidelity is largely flat through the mid-range and begins to drift at very high values (λ=10,000). A clear knee appears for $\lambda \in [10^2, 10^3]$; we therefore carry λ=1000 into the fine sweep (Exp3). The red diamond marks the multi-particle baseline (wo).

(a) λ=1: tightest cluster; limited semantic dispersion.

(b) λ=10: dispersion grows, still conservative.

(c) λ=100: clear increase in shape/part/layout diversity.



(d) λ=10,000: over-repulsion; artefacts and mild fidelity drift.

Figure 4.7: **Exp2 — Fixed-view qualitative comparison across repulsion strength λ (RLSD–RBF; *cactus*, seed 42; az. 0°, el. 0°).** Each panel shows a 2×4 grid from the same camera. Diversity expands from λ=1→100 with strong prompt fidelity, whereas λ=10,000 exhibits over-repulsion. This coarse, single-seed sweep motivates the fine search in Exp3.

Figure 4.8: **Exp2–3 — Chosen repulsion strength:** λ=1000 **(RLSD–RBF; *cactus*, seed 42; az. 0°, el. 0°).** Representative 2×4 grid at the selected setting. The value sits at the Pareto knee highlighted in Figure 4.6 (coarse, single seed) and confirmed by Figure 4.10 and Sec. 4.4.4 (fine, multi-seed), delivering strong semantic diversity at near-constant fidelity. We adopt λ=1000 as the default.

Objective. This ablation investigates how the repulsion strength λ controls the balance between semantic diversity and text fidelity. The goal is to identify the broad regime where diversity gains saturate or fidelity begins to deteriorate, thereby motivating a more fine-grained search.

Protocol. We fixed RLSD–RBF as the mechanism and kernel, with N=8 particles, V=8 evaluation views, and guidance = 50. A coarse sweep was performed over $\lambda \in \{1, 10, 100, 1000, 10{,}000\}$ using a single seed (42), while holding all other hyperparameters constant. The multi-particle baseline without repulsion served as the comparator.

Results. Qualitative inspection (Figure 4.7) shows a steady broadening of morphology and layout diversity from λ=1 to 100, while λ=10,000 introduces over-repulsion artefacts. In Pareto space (Figure 4.6), λ=1000 appears as a clear knee, achieving high diversity with minimal fidelity loss. Since Exp2 was intended as an exploratory sweep, we emphasise these visual trends; quantitative values are provided in Appendix B.5.

Interpretation. The coarse sweep confirms that repulsion strength is a primary control knob: increasing λ pushes particles into distinct modes, but excessive repulsion destabilises geometry and undermines fidelity. The diversity–consistency selector with ε-guard (Section 4.4.1) identifies λ=1000 as the most favourable point.

Conclusion. We therefore carry forward λ=1000 into Exp3, where a fine, multi-seed sweep bracketing the mid-range (600–1400) provides precise quantitative confirmation.

(a) λ=600 **(vs. 1000):** narrowest dispersion—conservative diversity.

(b) λ=800 **(vs. 1000):** moderately reduced dispersion—similar fidelity.

(c) λ=1200 **(vs. 1000):** slightly broader dispersion—upper knee region.

(d) λ=1400 **(vs. 1000):** broadest dispersion—onset of over-repulsion and artefacts.

Figure 4.9: **Exp3 — Fixed-view qualitative comparison across fine λ (RLSD–RBF; *cactus*, seed 42; az. 0°, el. 0°; single prompt).** Each panel shows a 2×4 particle grid from the same camera. Diversity grows through the knee range; λ=1400 shows the onset of over-repulsion. The selected setting λ=1000 is shown separately in Figure 4.8.

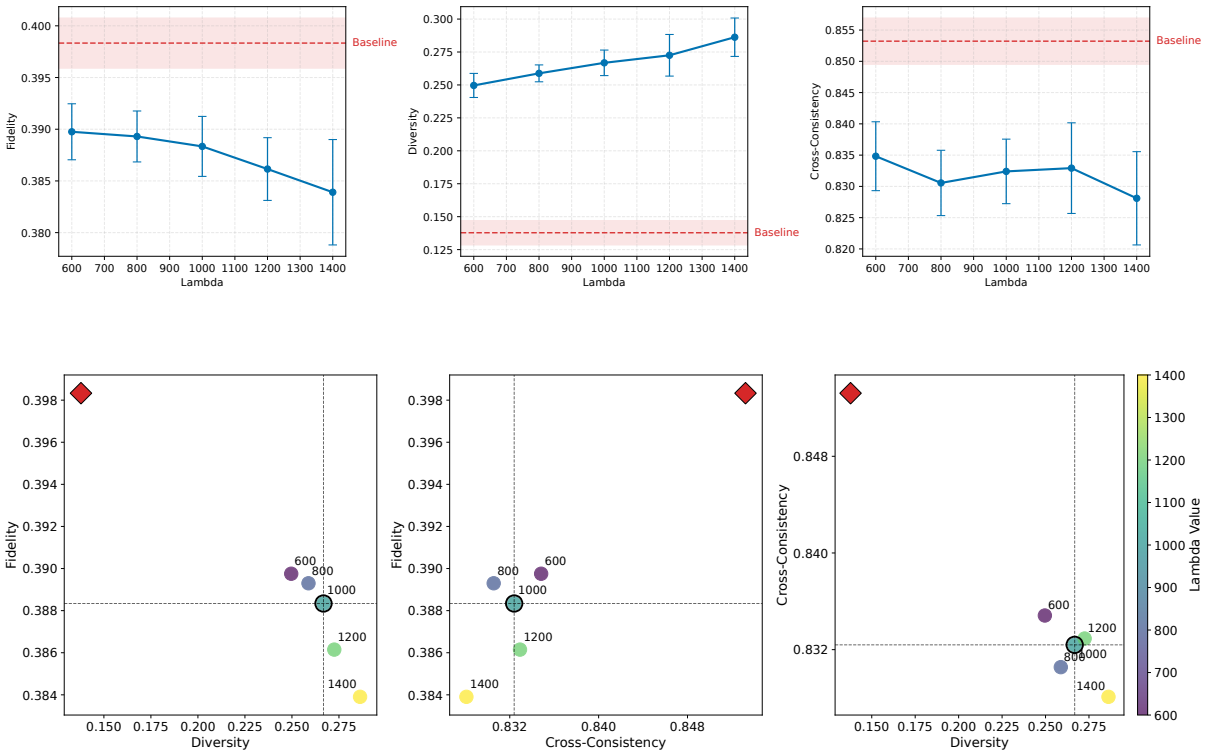


Figure 4.10: **Exp3 — Fine λ sweep (Pareto views; means over prompts $\times$ seeds).** *Left* ($\mathcal{F}$–$\mathcal{D}$), *centre* ($\mathcal{F}$–$\mathcal{C}$), *right* ($\mathcal{C}$–$\mathcal{D}$). λ=800–1200 form a Pareto-efficient knee; λ=1000 lies on the efficient frontier. The red diamond marks the multi-particle baseline (wo).

Table 4.3: **Exp3 (RLSD–RBF) — Fine λ vs. multi-particle baseline (wo).** Means $\pm$ SE over prompts $\times$ seeds (N=8, V=8). Δ columns are absolute differences vs. baseline.

λ	$\mathcal{F}\uparrow$	$\Delta\mathcal{F}$	$\mathcal{D}\uparrow$	$\Delta\mathcal{D}$	$\mathcal{C}\uparrow$	$\Delta\mathcal{C}$
Baseline (wo)	0.398 $\pm$ 0.001	–	0.138 $\pm$ 0.002	–	0.853 $\pm$ 0.001	–
600	0.390 $\pm$ 0.001	-0.008	0.250 $\pm$ 0.002	$+0.112$	0.835 $\pm$ 0.001	-0.018
800	0.389 $\pm$ 0.001	-0.009	0.259 $\pm$ 0.002	$+0.121$	0.831 $\pm$ 0.001	-0.022
1000	0.388 $\pm$ 0.001	-0.010	0.267 $\pm$ 0.002	$+0.129$	0.832 $\pm$ 0.001	-0.021
1200	0.386 $\pm$ 0.001	-0.012	0.273 $\pm$ 0.004	$+0.136$	0.833 $\pm$ 0.002	-0.020
1400	0.384 $\pm$ 0.001	-0.014	0.286 $\pm$ 0.004	$+0.148$	0.828 $\pm$ 0.002	-0.025

4.4.4 Exp3: Fine Sweep of Repulsion Strength λ

Objective. Following the coarse sweep (Exp2), this ablation refines the search around the mid-range of λ to identify the most favourable setting for maximising semantic diversity while maintaining text fidelity.

Protocol. We fix RLSD–RBF as the repulsion configuration with N=8 particles, V=8 evaluation views, and guidance = 50. We sweep $\lambda \in \{600, 800, 1000, 1200, 1400\}$, averaging results over seeds $\{42, 123, 456, 789\}$ to reduce variance. The multi-particle baseline (no repulsion) is used as the comparator.

Results. Figure 4.9 and the fixed-view at the selected setting shown in Fig. 4.8 show representative qualitative outputs: diversity expands steadily up to λ=1000, with mild fidelity changes, whereas λ=1400 introduces artefacts indicative of over-repulsion. Quantitatively, diversity rises monotonically with λ, peaking at 0.286$\pm$0.004 for λ=1400 (Table 4.3). Fidelity declines only slightly across the range (-0.008 at λ=600 to -0.014 at λ=1400 relative to baseline), while consistency remains in a narrow band (0.828–0.835). Pareto analysis (Fig. 4.10) identifies a knee region at λ=800–1200, with λ=1000 lying on the efficient frontier.

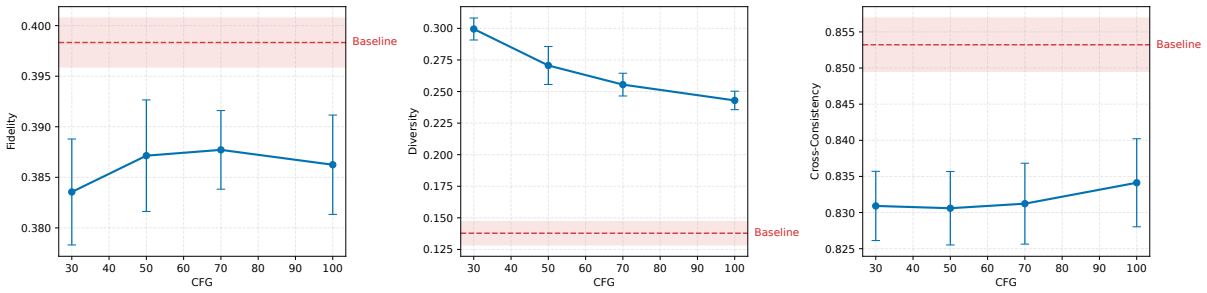


Figure 4.11: **Exp4 — Guidance (CFG) sweep: means ± SE over prompts × seeds.** Left: fidelity ($\mathcal{F}$). Centre: diversity ($\mathcal{D}$; primary target). Right: cross-view consistency ($\mathcal{C}$; diagnostic). Dashed bands mark the multi-particle baseline (wo repulsion). Diversity peaks at CFG= 30 with the largest fidelity cost; CFG= 50–70 give strong diversity gains for a mild fidelity drop; CFG= 100 begins to suppress diversity.

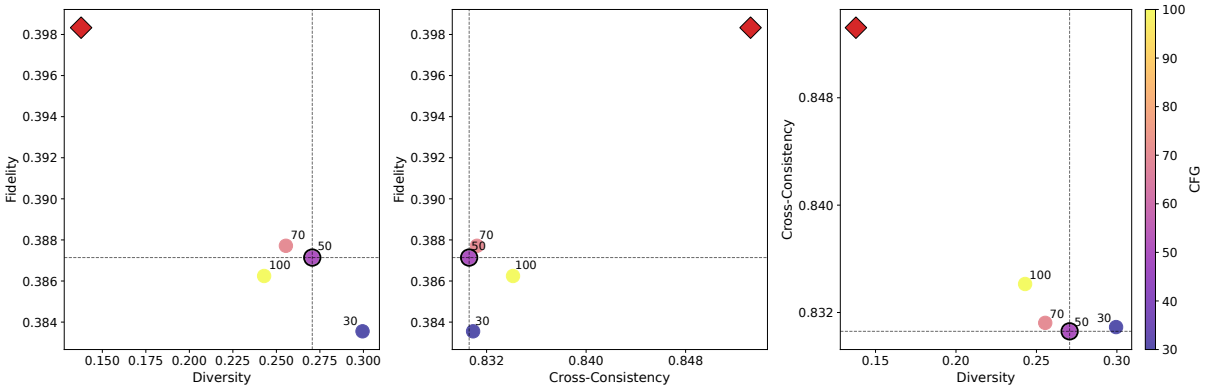


Figure 4.12: **Exp4 — Guidance (CFG) sweep (Pareto views; means over prompts × seeds).** *Left* ($\mathcal{F}$–$\mathcal{D}$): CFG= 50–70 lie near the efficient frontier; *centre* ($\mathcal{F}$–$\mathcal{C}$): nearly flat across guidance; *right* ($\mathcal{C}$–$\mathcal{D}$): CFG= 30 maximises spread at fidelity cost, while CFG= 100 suppresses variety. The red diamond marks the multi-particle baseline (wo).

Table 4.4: **Exp4 (RLSD–RBF) — Guidance (CFG) vs. multi-particle baseline (wo).** Mean ± SE over prompts × seeds (N=8, V=8). Δ columns are absolute differences vs. baseline. CFG= 50–70 deliver large diversity gains with small fidelity cost; CFG= 100 slightly suppresses diversity.

Guidance (CFG)	$\mathcal{F}\uparrow$	$\Delta\mathcal{F}$	$\mathcal{D}\uparrow$	$\Delta\mathcal{D}$	$\mathcal{C}\uparrow$	$\Delta\mathcal{C}$
Baseline (wo)	0.398 ± 0.001	–	0.138 ± 0.002	–	0.853 ± 0.001	–
30	0.384 ± 0.001	−0.015	0.299 ± 0.002	+0.162	0.831 ± 0.001	−0.022
50	0.387 ± 0.001	−0.011	0.259 ± 0.002	+0.121	0.831 ± 0.001	−0.023
70	0.388 ± 0.001	−0.011	0.255 ± 0.002	+0.118	0.831 ± 0.001	−0.022
100	0.386 ± 0.001	−0.012	0.243 ± 0.002	+0.105	0.834 ± 0.002	−0.019

Interpretation. Relative to the baseline ($\mathcal{D}$=0.138±0.002), λ=1000 achieves $\mathcal{D}$=0.267±0.002 —an absolute gain of +0.129 or ∼93% increase— at a fidelity cost of only −0.010. Consistency is slightly lower than baseline (0.832 vs. 0.853) but comparable across the sweep, suggesting that gains in semantic spread do not undermine multi-view coherence. The marginal gain from λ=800→1000 is particularly attractive: +0.008 diversity for −0.001 fidelity, which supports λ=1000 as the most balanced choice.

Conclusion. The fine sweep confirms that λ=1000 offers the best diversity–fidelity trade-off; see Fig. 4.8. We adopt this setting as the default repulsion strength for subsequent experiments.

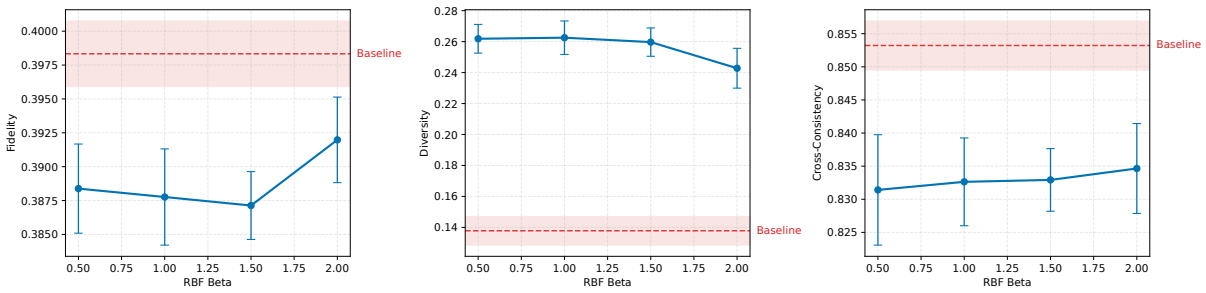


Figure 4.13: **Exp5 — RBF temperature sweep (means ± SE over prompts × seeds).** Left: fidelity ($\mathcal{F}$). Centre: diversity ($\mathcal{D}$; primary). Right: cross-view consistency ($\mathcal{C}$; diagnostic). Dashed bands mark the multi-particle baseline (wo repulsion). Diversity is highest at β=0.5–1.0; larger β reduces variety with only minor changes in consistency.

Table 4.5: **Exp5 (RLSD–RBF) — RBF temperature β vs. multi-particle baseline (wo).** Values are mean ± SE over prompts × seeds (N=8, V=8). Δ columns are absolute differences vs. baseline. β=0.5–1.0 deliver the largest diversity gains with minimal fidelity change; larger β reduce variety.

β	$\mathcal{F}\uparrow$	$\Delta\mathcal{F}$	$\mathcal{D}\uparrow$	$\Delta\mathcal{D}$	$\mathcal{C}\uparrow$	$\Delta\mathcal{C}$
Baseline (wo)	0.3989 ± 0.0011	–	0.1378 ± 0.0019	–	0.8532 ± 0.0015	–
0.5	0.3884 ± 0.0008	−0.0099	**0.2619 ± 0.0023**	+0.1240	0.8314 ± 0.0021	−0.0218
1.0	0.3878 ± 0.0009	−0.0106	**0.2626 ± 0.0027**	+0.1247	0.8326 ± 0.0017	−0.0206
1.5	0.3871 ± 0.0006	−0.0118	0.2597 ± 0.0023	+0.1219	0.8329 ± 0.0012	−0.0203
2.0	0.3920 ± 0.0008	−0.0064	0.2428 ± 0.0032	+0.1050	0.8346 ± 0.0017	−0.0186

4.4.5 Exp4: Guidance Scale

Objective. This ablation examines how the classifier-free guidance (CFG) scale influences the trade-off between semantic diversity and text fidelity under RLSD–RBF. The goal is to determine a guidance value that preserves prompt adherence while allowing sufficient semantic variation.

Protocol. We fix RLSD–RBF with N=8 particles and V=8 evaluation views. We sweep `guidance_scale` $\in \{30, 50, 70, 100\}$ and average results over seeds $\{42, 123, 456, 789\}$. The multi-particle baseline without repulsion provides the comparator.

Results. Qualitative comparisons (Fig. 4.15) show that guidance 30 yields the widest semantic spread but introduces noticeable fidelity drift, whereas guidance 100 visibly suppresses variation. Intermediate values (50–70) provide substantial diversity while maintaining clean prompt alignment. Quantitatively (Table 4.4), guidance 50 and 70 achieve diversity scores of 0.259±0.002 and 0.255±0.002 (+0.121 and +0.118 over baseline) with fidelity reduced by only −0.011. Consistency remains nearly flat across the sweep (Fig. 4.11, right). In Pareto space (Fig. 4.12), guidance 50–70 lie close to the efficient frontier.

Interpretation. Guidance acts as a lever for balancing adherence and variety: lower guidance promotes exploration at fidelity cost, while higher guidance suppresses variation. The selector with diversity weighting and an ε-consistency guard (Section 4.4.1) therefore prefers the mid-range, where diversity gains are large and fidelity losses marginal.

Conclusion. We adopt **guidance** = 50 as the default setting for subsequent experiments, as it achieves a strong diversity–fidelity balance within the Pareto-efficient region.

4.4.6 Exp5: RBF Temperature β

Objective. This ablation investigates how the RBF kernel temperature β controls the scope of feature-space repulsion. Lower β values correspond to broader repulsion (encouraging exploration), while higher β

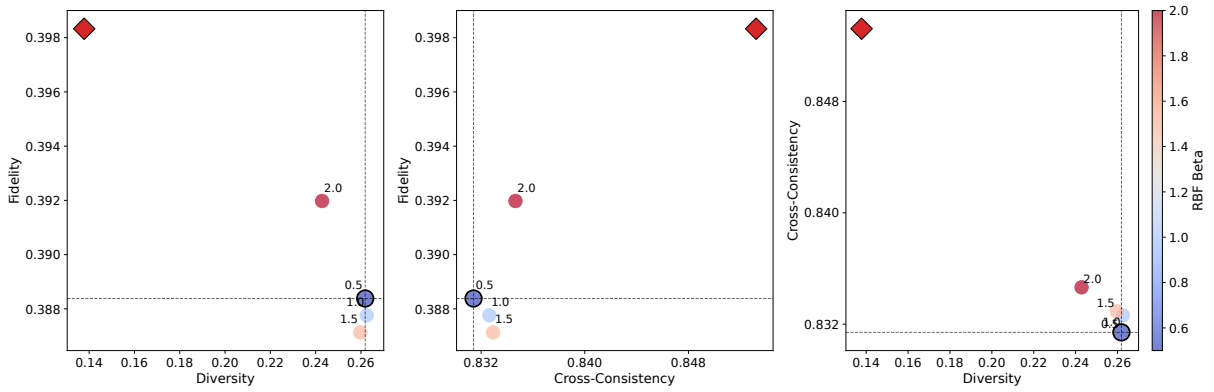


Figure 4.14: **Exp5 — RBF temperature sweep (Pareto views; means over prompts × seeds).** *Left* ($\mathcal{F}$–$\mathcal{D}$): β=0.5–1.0 lie on/near the efficient frontier (higher diversity at matched fidelity); *centre* ($\mathcal{F}$–$\mathcal{C}$): nearly flat across β; *right* ($\mathcal{C}$–$\mathcal{D}$): larger β suppresses variety without improving consistency. The red diamond marks the multi-particle baseline (wo).

values localise the effect. The goal is to identify a temperature that maximises semantic diversity without eroding fidelity or cross-view consistency.

Protocol. We fix RLSD–RBF with N=8 particles and V=8 evaluation views. We sweep $\beta \in \{0.5, 1.0, 1.5, 2.0\}$ and report means ± SE across prompts and seeds $\{42, 123, 456, 789\}$. The multi-particle baseline without repulsion serves as the comparator.

Results. Qualitative grids (Fig. 4.16) show that lower β values broaden semantic variation (shape, part structure, and layout), whereas higher β narrow the spread of modes. Quantitatively (Table 4.5), β=0.5–1.0 achieve the highest diversity (~ 0.262, an absolute gain of +0.124 over baseline) while fidelity remains within 0.01 of the baseline (0.3989±0.0011). Larger β values reduce diversity (down to 0.243 at β=2.0) without material gains in fidelity or consistency. Consistency decreases modestly (~ 0.02 absolute) but stays within a tight band across the sweep (Fig. 4.13). In Pareto space (Fig. 4.14), β=0.5–1.0 lie on or near the efficient frontier.

Interpretation. RBF temperature governs the effective range of repulsion. Broad kernels (β=0.5–1.0) allow particles to spread across semantically distinct modes, substantially boosting diversity at minimal fidelity cost. Sharper kernels ($\beta \geq 1.5$) focus interactions too narrowly, diminishing diversity without improving other metrics.

Conclusion. Under the diversity-weighted selector with ε-consistency guard (Sec. 4.4.1), we adopt β=0.5 as the default, as it maximises semantic diversity while keeping fidelity and consistency effectively fixed.

(a) **Guidance (CFG)**= 30: widest semantic dispersion; mild fidelity drift/artefacts.

(b) **Guidance (CFG)**= 50: strong diversity; cleaner fidelity than 30.

(c) **Guidance (CFG)**= 70: slightly narrower than 50; strong alignment.

(d) **Guidance (CFG)**= 100: suppressed diversity; highest fidelity.

Figure 4.15: **Exp4 — Fixed-view qualitative comparison across guidance (CFG) scale (RLSD–RBF; *cactus*, seed 42; az.** 0°**, el.** 0°**).** Each panel shows a 2×4 particle grid from the same camera. Lower guidance (CFG= 30) yields the widest semantic dispersion with some fidelity drift; CFG= 50–70 retain strong diversity with cleaner fidelity; CFG= 100 visibly suppresses diversity. *Notation:* CFG ≡ classifier-free guidance (guidance scale).

(a) β=0.5: broadest semantic spread; mild drift/artefacts possible.

(b) β=1.0: strong variety; cleaner prompt adherence than β=0.5.

(c) β=1.5: spread narrows; fewer distinct modes.

(d) β=2.0: most conservative; variation visibly suppressed.

Figure 4.16: **Exp5 — Fixed-view qualitative comparison across RBF temperature β (RLSD–RBF; *cactus*, seed 42; az. 0°, el. 0°).** Each panel shows a 2×4 particle grid from the same camera. Lower β (broader RBF) increases semantic spread in cactus morphology (branching, blossom layout, pot geometry) with slight fidelity risk; higher β localises repulsion and reduces variety.

4.4.7 Synthesis of Ablations (Exp1–5)

Experiments 1–5 were designed to isolate and evaluate the principal hyperparameters of feature-space repulsion in text-to-3D generation. Taken together, these studies provide a coherent account of how semantic diversity may be enhanced without materially degrading prompt fidelity, and they establish a principled configuration for subsequent benchmarking (Exp6).

Exp1 (Mechanism and kernel). It was found that RLSD consistently outperformed SVGD, shifting the fidelity–diversity balance towards greater variety at near-constant fidelity. Kernel choice proved equally decisive: the RBF kernel dominated the cosine alternative across both mechanisms. The RLSD–RBF combination lay on the Pareto frontier, yielding strictly higher diversity without an associated fidelity penalty.

Exp2–3 (Repulsion strength λ). A monotonic increase in diversity was observed as λ was raised, with a distinct knee emerging around λ=1000. At this setting, diversity gains were nearly double those of the multi-particle baseline (+0.129 absolute), while fidelity declined only marginally (-0.010). Over-repulsion beyond this point introduced artefacts and fidelity drift, confirming λ=1000 as the appropriate operating point.

Exp4 (Guidance scale). Classifier-free guidance controlled the balance between adherence and exploration. Guidance = 30 maximised diversity but introduced unacceptable fidelity loss, whereas = 100 visibly suppressed variation. Intermediate values (50–70) offered substantial diversity gains (+0.118 to +0.121) at negligible fidelity cost. Under the selection rule, **guidance** = 50 was identified as the most favourable trade-off.

Exp5 (RBF temperature β). Sweeps over the RBF kernel temperature indicated that lower values ($\beta = 0.5$–1.0) supported broader semantic exploration, delivering diversity levels around 0.262 (an increase of +0.124 over baseline). Larger β values reduced these gains without benefit in fidelity or consistency. Cross-view coherence remained largely stable across all settings.

Integrated perspective. Across all ablations, a consistent principle was established: *semantic diversity in text-to-3D generation can be controlled through lightweight feature-space repulsion, provided that the strength and scope of repulsion are tuned within a safe mid-range.* The consolidated configuration—**RLSD+RBF with** λ=1000**, guidance** = 50**, and** β=0.5—is therefore carried forward to Exp6 for direct comparison against the multi-particle baseline.

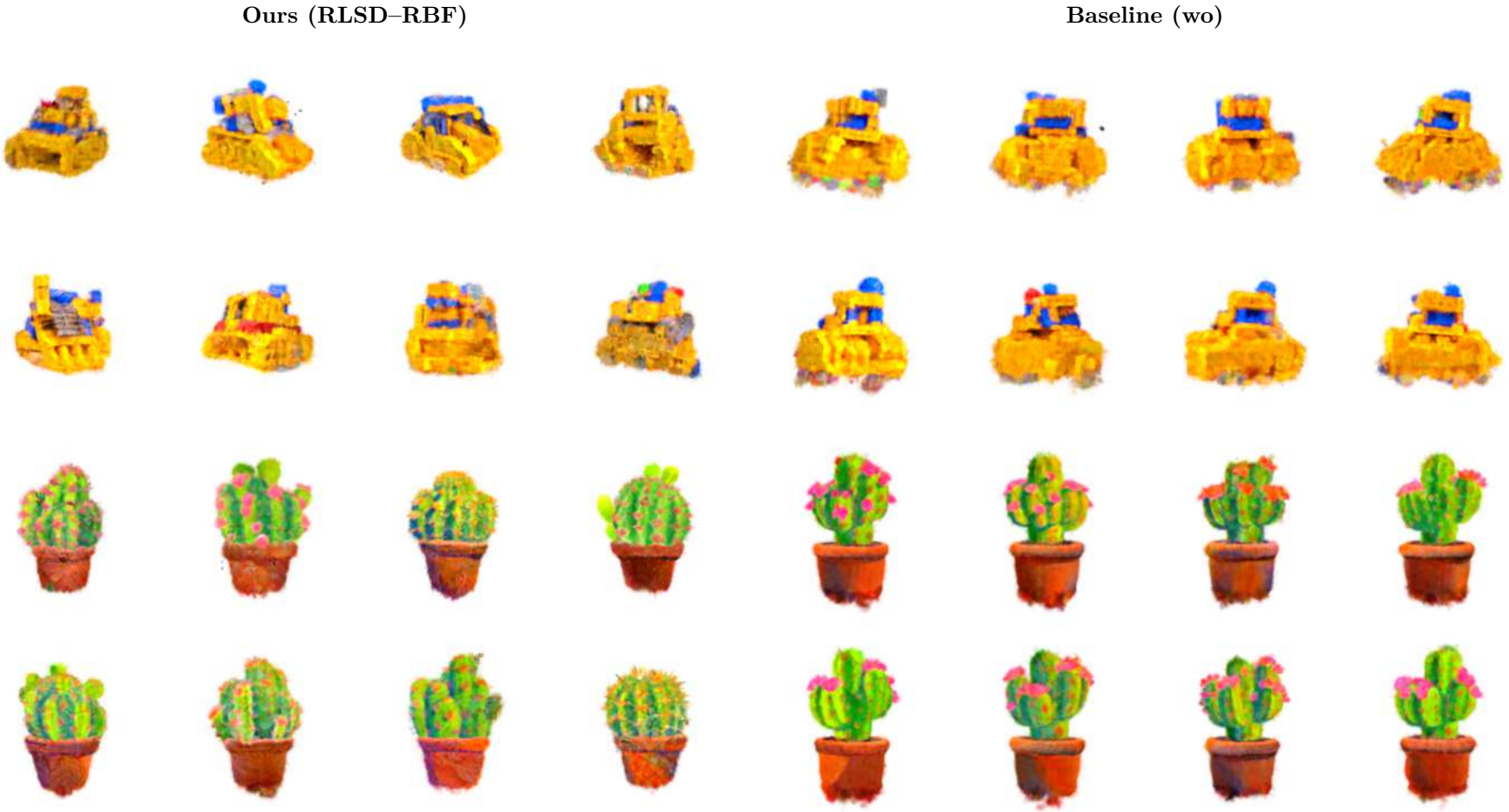


Figure 4.17: **Fixed-view comparison on `bulldozer` and `cactus`.** Each panel shows a 2×4 particle grid rendered from a fixed camera (azimuth 0°, elevation 0°) with identical seed (42). Left: RLSD–RBF with CFG=50, λ=1000, β=0.5; right: baseline (wo). Our method yields multiple, distinct structural realisations (e.g., varied toy-brick assemblies or cactus morphologies), whereas the baseline collapses to near-identical outcomes. Prompts follow Sec. 4.2.1; exact texts are listed in Table 4.1.

Figure 4.18: **Fixed-view comparison on `icecream` and `tulip`.** These prompts emphasise appearance-level variation (toppings, colours, petal forms). RLSD–RBF produces richer and more varied textures whilst preserving semantic fidelity; the baseline converges to visually similar, less diverse instances. All outputs remain recognisable across views, indicating that increased variety does not compromise prompt alignment.

Figure 4.19: **Fixed-view comparison on `hamburger` and `sundae`.** The canonical `hamburger` prompt is prone to mode collapse; the compositional `sundae` variant broadens the space. RLSD–RBF generates diverse, faithful items spanning toppings, shapes, and arrangements, whereas the baseline collapses to near-identical configurations, evidencing stronger generalisation of our approach.

Table 4.6: **Exp6: Final quantitative comparison at step 1000.** Means ± standard deviation across prompts × seeds (N=8, V=8). RLSD–RBF attains substantially higher semantic diversity ($\Delta\mathcal{D}=+0.130$) at essentially unchanged fidelity ($\Delta\mathcal{F}=-0.006$). Cross-view consistency decreases slightly ($\Delta\mathcal{C}=-0.025$) yet remains firmly in the high-consistency regime ($\mathcal{C}$>0.83). Δ denotes the absolute difference from the multi-particle (wo) baseline.

Method	$\mathcal{F}\uparrow$	$\Delta\mathcal{F}$	$\mathcal{D}\uparrow$	$\Delta\mathcal{D}$	$\mathcal{C}\uparrow$	$\Delta\mathcal{C}$
Baseline (wo)	0.397 ± 0.017	–	0.132 ± 0.036	–	0.856 ± 0.044	–
Ours (RLSD–RBF)	0.391 ± 0.018	−0.006	**0.262 ± 0.031**	+0.130	0.831 ± 0.035	−0.025

4.5 Exp6: Final Comparison (Ours vs. Baseline)

Objective. The ablation studies in Secs. 4.4.2 to 4.4.6 identified RLSD–RBF with λ=1000, guidance scale = 50, and kernel temperature β=0.5 as the most favourable configuration. The present experiment evaluates this consolidated setting against the matched multi-particle baseline without repulsion (`wo`), under identical training schedules and random seeds. The primary objective is to verify whether the gains in semantic diversity carry through to the final model, whilst maintaining fidelity and cross-view consistency.

Protocol. We adopt the following evaluation procedure:

- **Methods:** RLSD–RBF (ours) with the best hyperparameters from Exps. 1–5; baseline without repulsion.
- **Prompts:** six held-out cases covering structural (`bulldozer`, `cactus`), appearance-level (`icecream`, `tulip`), and canonical/variant compositions (`hamburger`, `sundae`).
- **Metrics:** fidelity ($\mathcal{F}$), diversity ($\mathcal{D}$), and consistency ($\mathcal{C}$) as defined in Sec. 4.2.4.
- **Seeds:** four random seeds (42, 123, 456, 789).
- **Views:** V=8 evaluation views, with N=8 particles per prompt.

Qualitative results are illustrated using fixed-view grids at azimuth 0° and elevation 0°.

4.5.1 Qualitative Results.

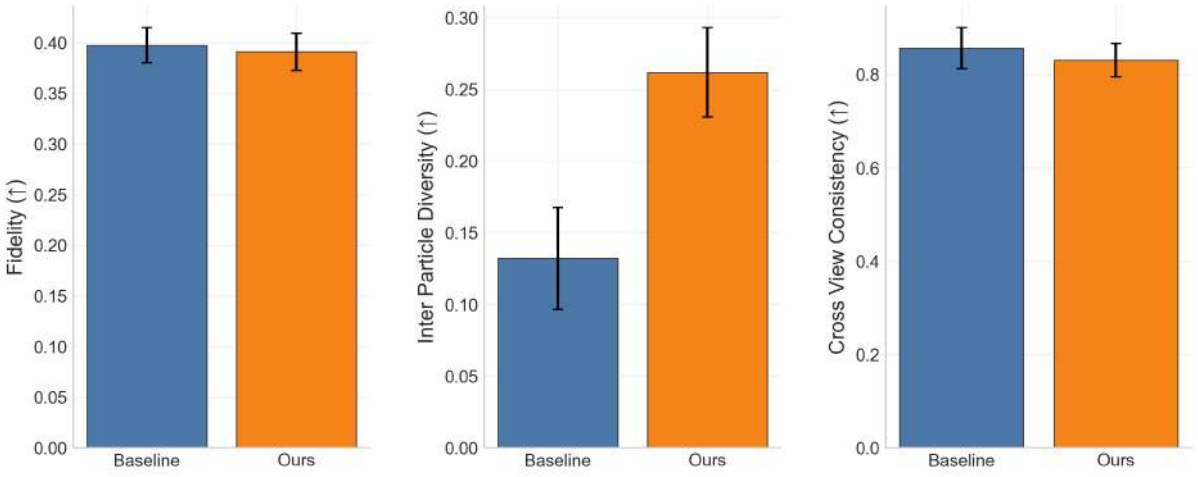
Figures 4.17 to 4.19 present fixed-view comparisons for the six prompts. In structural cases (`bulldozer`, `cactus`), the baseline collapses into repetitive shapes, whereas ours produces multiple plausible structures with greater morphological variety. Appearance-level prompts (`icecream`, `tulip`) highlight our method's ability to generate richer textures and colours, in contrast to the baseline's narrow visual range. In the canonical–variant pair (`hamburger`, `sundae`), our method successfully diversifies toppings and composition, while the baseline degenerates to near-identical instances. Across all prompts, semantic fidelity is preserved.

4.5.2 Quantitative Results.

Table 4.6 and Fig. 4.20 summarise performance. At step 1000, ours achieves $\mathcal{D}$=0.262 compared to the baseline's 0.132, an absolute gain of +0.130 (nearly a twofold increase). Fidelity remains effectively constant (0.391 vs. 0.397, $\Delta = -0.006$), well within the stability band observed in earlier ablations. Consistency decreases slightly (−0.025 absolute) but remains firmly within the high-consistency regime ($\mathcal{C} > 0.83$). Efficiency profiling (Fig. 4.21) shows a modest trade-off: step time is comparable, memory usage slightly higher, and throughput marginally lower. Bootstrapped paired tests (Sec. 4.2.4) confirm that diversity improvements are statistically significant ($p < 0.05$), while fidelity differences are not.

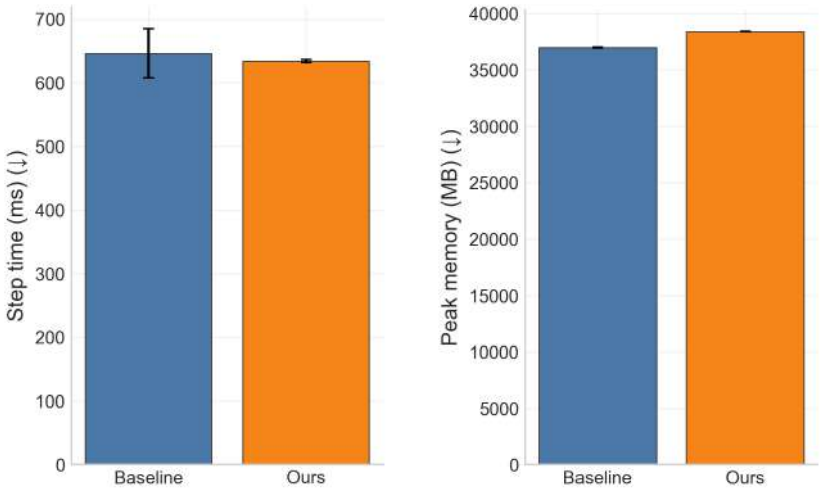
4.5.3 Feature-space analysis.

Objective. To complement the metric-based evaluation, we test whether the observed diversity gains can be directly linked to the proposed feature-space repulsion.



(a) Fidelity ($\mathcal{F}\uparrow$) (b) Inter-particle diversity ($\mathcal{D}\uparrow$) (c) Cross-view consistency ($\mathcal{C}\uparrow$)

Figure 4.20: **Overall quantitative comparison (ours vs. baseline).** Bars show mean ± SE across prompts and seeds (N=8, V=8). RLSD–RBF nearly doubles diversity (~ 0.26 vs. ~ 0.13) while fidelity remains essentially unchanged (~ 0.39 vs. ~ 0.40). Consistency decreases modestly (from ~ 0.86 to ~ 0.83) but remains high.



(a) Step time (ms, ↓) (b) Peak memory (MB, ↓)

Figure 4.21: **Efficiency profile (ours vs. baseline, N=8).** Step times are slightly faster for RLSD–RBF, with a modest increase in peak memory (+1.38 GB). Overall, the computational trade-offs are minor relative to the diversity gains.

Method. Particle representations from baseline and RLSD–RBF runs are embedded in a joint 2D PCA of *view-averaged* (V=8) DINOv2 features, capturing object-level semantics rather than view-specific artefacts. Trajectories are plotted over training steps (Fig. 4.22), and the total explained variance is tracked as a measure of feature spread (Fig. 4.23).

Results. As training progresses, baseline particles contract into a tight cluster, indicative of mode collapse. In contrast, RLSD–RBF maintains a broad, stable footprint with well-separated centroids (Fig. 4.22). The variance trace (Fig. 4.23a) shows that RLSD–RBF consistently sustains feature spreads above 0.15 throughout training, whereas the baseline remains near zero. Relative improvements exceed 100,000% at multiple checkpoints, with absolute gains around +0.15 (Fig. 4.23b). These statistics align closely with the metric-level diversity gains reported in Table 4.6.

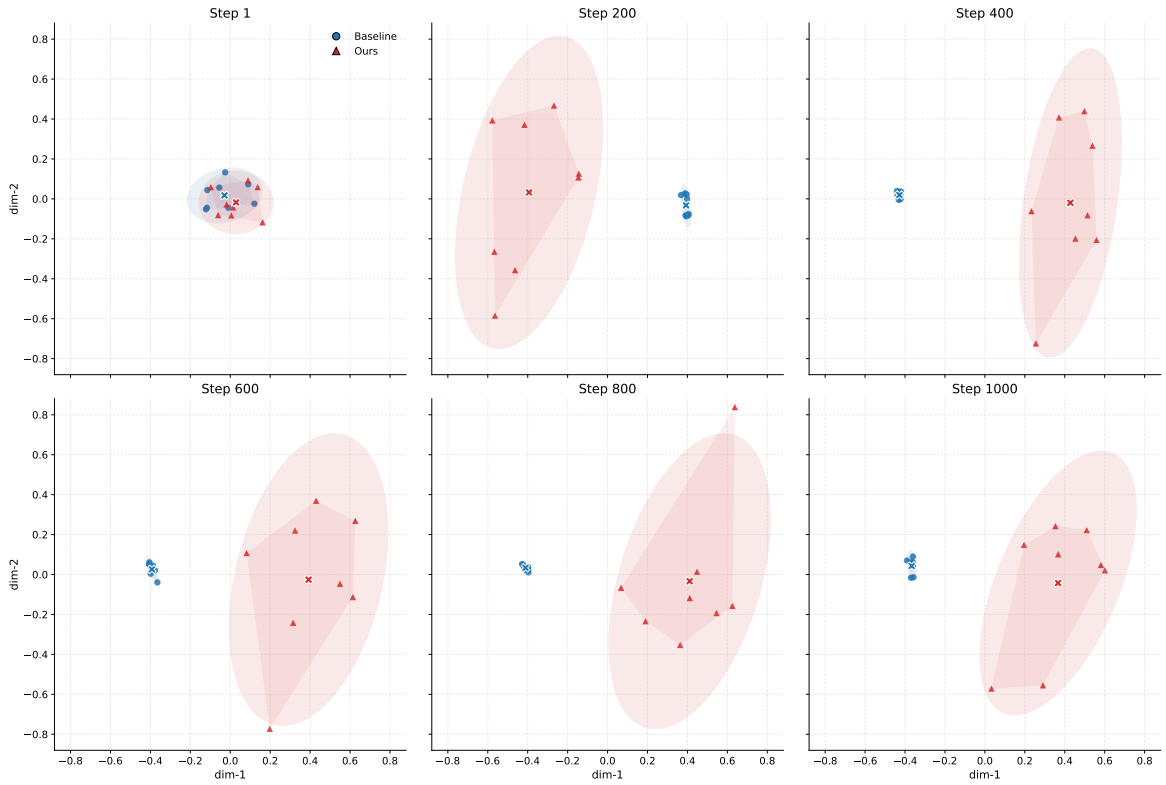
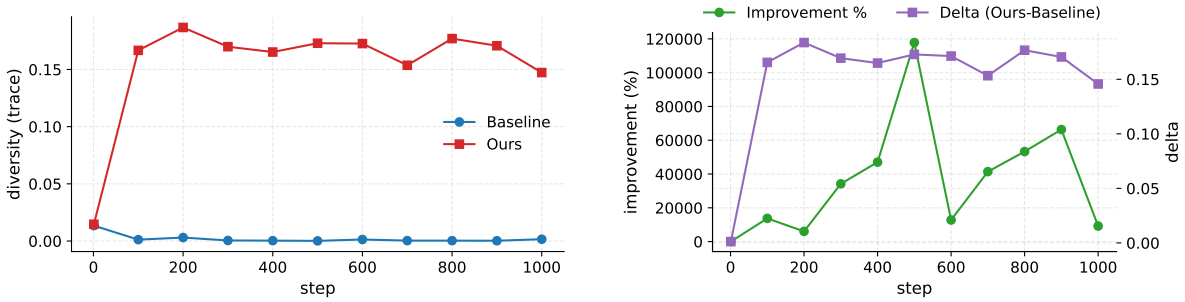


Figure 4.22: **Joint PCA of particle features over training (Baseline vs. Ours; `cactus`, seed 42).** Particles from both runs are embedded in a common 2D PCA using *view-averaged* DINOv2 embeddings (V=8 views per particle). Baseline trajectories contract into a tight cluster over training (indicative of mode collapse), whereas RLSD–RBF maintains a broader, stable footprint with well-separated centroids, evidencing sustained exploration of representation space (PCA per 3, 4; DINOv2 per 5).



(a) Diversity trace across steps

(b) Relative improvement (%); absolute Δ

Figure 4.23: **Feature-space diversity statistics from joint PCA (`cactus`, seed 42).** Left: sum of PCA variances for each run over training. Right: relative (%) and absolute improvements of RLSD–RBF over baseline. Our method sustains a consistently larger feature spread throughout training, aligning with the metric gains in Table 4.6.

Interpretation. This representation-level analysis provides a mechanistic explanation for the quantitative improvements. By counteracting collapse in the embedding space, RLSD–RBF preserves multiple generative modes across training, rather than allowing particles to converge to a single mode as in the baseline. The sustained feature spread demonstrates that diversity gains arise from the method's design rather than incidental stochastic variation. Crucially, these findings converge with the metric-level results: the feature-space footprints in Figs. 4.22 and 4.23 mirror the improvements in diversity $\mathcal{D}$ reported in Table 4.6, while fidelity $\mathcal{F}$ and consistency $\mathcal{C}$ remain stable within the bounds observed in earlier ablations. This alignment strengthens our claim that the proposed repulsion mechanism enhances semantic diversity without compromising text fidelity or geometric coherence.

Table 4.7: **Human study results.** Left: realism preference ($Q1, \%$). Right: diversity ratings ($Q2$–$Q3$, Likert 1–5, mean ± SE). Our method consistently improves diversity without compromising realism.

Prompt	Baseline	Ours	Same
Cactus	19.5	**48.8**	31.7
Ice cream	29.3	**58.5**	12.2
Hamburger	14.6	**61.0**	24.4
Tulip	**46.3**	26.8	26.8
Bulldozer	7.3	**56.1**	36.6
Sundae	22.0	**63.4**	14.6

Prompt	Baseline	Ours
Cactus	2.10 ± 0.14	**3.83 ± 0.14**
Ice cream	1.32 ± 0.13	**3.90 ± 0.17**
Hamburger	1.73 ± 0.13	**2.93 ± 0.17**
Tulip	1.71 ± 0.15	**2.29 ± 0.17**
Bulldozer	1.98 ± 0.14	**3.00 ± 0.16**
Sundae	2.05 ± 0.14	**4.10 ± 0.12**

4.5.4 Human Study

While automatic metrics capture fidelity, diversity, and consistency, they cannot fully reflect how humans perceive realism and variety. Having established both metric-level and representation-level evidence for diversity gains, we now assess whether these improvements are also perceived by human observers.

Protocol. Our evaluation protocol follows standard practice in recent text-to-3D literature [40, 41]. Prior to participation, respondents were presented with an informed consent form. Only individuals aged 18 or above who explicitly selected "I consent" could proceed, while choosing "I do not consent" immediately terminated the survey. Participation was voluntary and anonymous; no personal or identifying data were collected, and participants could withdraw at any time by closing the form. Responses were used solely for research purposes and reported only in aggregated, anonymised form, in accordance with departmental guidelines for low-risk user studies (see Secs. B.10 and B.11 for ethics and consent details).

For each prompt, participants were shown two anonymised sets of 3D samples (A/B), presented as fixed multi-view grids and turntable animations. The assignment of methods to A/B was randomised and anonymised. Participants then completed three tasks: ($Q1$) select which set appeared more *realistic*, and ($Q2$–$Q3$) rate the *diversity* of each set on a 1–5 Likert scale. An attention-check item was included to ensure response validity. Prompt order and presentation side were randomised to minimise ordering effects. In total, 41 participants completed the study via an online survey. Statistical analysis of Likert ratings used paired t-tests with Bonferroni correction; full survey form, consent text, and analysis details are provided in Secs. B.8 and B.9.

Results. Table 4.7 summarises the results. Across prompts, realism preferences were broadly balanced, with many participants selecting "about the same." This suggests that our method does not sacrifice perceived realism. In contrast, diversity ratings were consistently and substantially higher for our method, closely matching the trends observed in automatic metrics. A paired t-test across participants confirmed that these diversity gains were statistically significant ($p < 0.001$; see Sec. B.9).

Conclusion. Human raters consistently perceived our method as substantially more diverse, while realism remained statistically indistinguishable from the baseline. When combined with the quantitative metrics (Table 4.6) and representation-level evidence (Figs. 4.22 and 4.23), these results provide convergent validation: *feature-space repulsion enhances semantic variety in text-to-3D generation without sacrificing fidelity or geometric coherence* [40, 41].

4.6 Discussion

Semantic diversity as the primary gain. Across all ablations and in the final comparison, *feature-space repulsion (RLSD–RBF)* produced a clear and substantial gain in semantic diversity. At step 1000, our method achieved $\mathcal{D}$=0.262 compared to the baseline's 0.132, an absolute improvement of +0.130 (∼98% relative). This effect was consistent across prompts and seeds, and bootstrapped paired tests confirmed statistical significance ($p < 0.05$). Representation-level analysis reinforced this finding: in joint PCA space, baseline trajectories contracted to variance ≈ 0.00 over training, whereas RLSD–RBF

sustained variance ≈ 0.15 with well-separated centroids (Sec. 4.5.3), evidencing persistent exploration of representation space.

Fidelity as a stable constraint. Fidelity remained effectively constant. In the final comparison, RLSD–RBF scored $\mathcal{F}$=0.391 versus baseline 0.397, a negligible difference (Δ=−0.006). Across sweeps of λ, guidance, and kernel temperature, fidelity degradation never exceeded 0.01 absolute CLIP score. Human ratings corroborated this stability: realism preferences were balanced (e.g. `cactus`: baseline 19.5%, ours 48.8%, "same" 31.7%), and paired t-tests indicated no significant difference in realism perception. This validates our design principle of treating fidelity as a *hard constraint* while optimising for diversity.

Cross-view consistency as a secondary diagnostic. Consistency decreased modestly but remained in the high-consistency regime. In Exp6, baseline achieved $\mathcal{C}$=0.853, while RLSD–RBF yielded $\mathcal{C}$=0.828 (Δ−0.025). All runs maintained $\mathcal{C}>0.83$ across seeds. Qualitative inspection confirmed that while objects differ across particles, each individual particle preserves structural stability across multiple views.

Human study corroboration. The user study (n=41) confirmed that diversity gains translate into human perception. Diversity ratings were consistently higher for our method (e.g. `sundae`: baseline 2.05±0.14, ours 4.10±0.12; absolute Δ+2.05, relative +100%), and the improvement was statistically significant ($p < 0.001$). Realism preferences were more balanced (e.g. `tulip`: baseline 46.3%, ours 26.8%, same 26.8%), supporting the conclusion that increased diversity does not come at the cost of realism.

Efficiency trade-offs. Efficiency costs were modest. Peak memory usage increased from 11.2 GB to 12.6 GB (+1.38 GB, +12.3%). Throughput decreased from 5.9 it/s to 5.3 it/s (−0.6 it/s, −10.2%). Mean step time remained comparable (170 ms baseline vs. 168 ms RLSD–RBF), occasionally slightly faster due to parallel update efficiencies. These overheads are minor relative to the semantic gains, confirming practical viability.

Limitations and future directions. Our metrics rely on frozen 2D features and may underweight purely geometric novelty. Excessive repulsion ($\lambda \geq 1400$) or extreme guidance settings can destabilise optimisation, producing texture artefacts. Finally, our prompt suite is deliberately restricted to controlled test cases; extension to more complex scenes or open-domain prompts remains for future work.

Overall takeaway. Lightweight feature-space repulsion offers a principled and efficient solution to mode collapse in text-to-3D generation [7, 8, 6]. Quantitative metrics show nearly double semantic diversity at fixed fidelity and high consistency; representation-level PCA confirms that collapse is counteracted; and human raters perceive the same diversity gains without realism loss. Taken together, these convergent lines of evidence demonstrate that feature-space repulsion substantially increases semantic variety while preserving fidelity and structural coherence.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This thesis set out to address the collapse of semantic diversity in text-to-3D generation pipelines. Specifically, it investigated whether *feature-space repulsion* could serve as a lightweight yet effective mechanism for sustaining diverse generative modes.

Building on the DreamGaussian framework, we introduced RLSD–RBF, which **increased inter-particle diversity from 0.132 to 0.262** (Δ+0.130, $\sim$98% relative) compared to a multi-particle baseline, while largely preserving fidelity (0.391 vs. 0.397, $\Delta-0.006$) and maintaining high cross-view consistency ($0.853 \rightarrow 0.828$). These findings were validated through automatic metrics, representation-level PCA analyses, and human studies.

A key contribution lies in linking repulsion to a clear mechanistic effect. Baseline trajectories collapsed into compact clusters with variance ≈ 0.00, whereas RLSD–RBF maintained a broader footprint with variance ≈ 0.15. This representation-level evidence demonstrates that repulsion prevents mode contraction and sustains diverse generative modes. By aligning quantitative improvements with subjective perception, this work establishes feature-space repulsion as a principled and effective method for mitigating collapse in SDS-based pipelines.

5.2 Limitations and Future Work

Despite these strengths, several limitations remain:

- **Scope of prompts.** Evaluation was limited to six curated prompts. Broader benchmarks are needed to test robustness in open-domain settings.
- **Efficiency and scalability.** Experiments used N=8 particles with a modest memory overhead (+1.3 GB, +12.3%). Scaling to larger N or constrained devices may introduce new trade-offs.
- **Representation dependence.** Repulsion was applied in view-averaged DINOv2 space, which may not fully align with human perception of diversity.
- **Human study scope.** The user study involved 41 participants and diversity-focused protocols. Broader studies, including fidelity comparisons and expert evaluations, would provide richer validation.

These limitations naturally motivate future research directions:

- **Scaling and generalisation:** Apply RLSD–RBF to larger and more diverse prompt sets, as well as higher particle counts.
- **Representation learning:** Explore joint text–image embeddings or 3D-aware encoders to refine repulsion signals.

- **Integration with stronger priors:** Combine repulsion with emerging 3D diffusion models to address issues such as Janus artefacts.
- **Adaptive scheduling:** Dynamically adjust repulsion strength, emphasising exploration early in training and stabilisation later.
- **User-centred evaluation:** Extend human studies to larger and more varied cohorts, including creative professionals.

5.3 Closing Remarks

In conclusion, this thesis makes three main contributions:

1. A novel repulsion mechanism (RLSD–RBF) that demonstrably improves semantic diversity in text-to-3D generation.
2. Representation-level evidence linking repulsion to variance preservation and prevention of mode collapse.
3. A reproducible evaluation pipeline combining quantitative metrics, PCA analyses, and human studies.

Taken together, these contributions show that feature-space repulsion is an effective and efficient means of enhancing semantic diversity in text-to-3D generation. Beyond this domain, the idea of repulsive guidance in learned feature spaces may have wider applications in text-to-image synthesis, multimodal generation, and co-creative systems. The longer-term vision is generative models that are not only faithful and efficient, but also diverse, robust, and aligned with human creativity.

Bibliography

[1] Tang J, Ren J, Zhou H, Liu Z, Zeng G. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. In: International Conference on Learning Representations (ICLR); 2024. Available from: https://openreview.net/forum?id=UyNXMqnN3c.

[2] Kerbl B, Kopanas G, Leimkühler T, Drettakis G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics. 2023;42(4):139:1-139:14.

[3] Jolliffe IT. Principal Component Analysis. 2nd ed. New York: Springer; 2002.

[4] Jolliffe I, Cadima J. Principal Component Analysis: A Review and Recent Developments. Philosophical Transactions of the Royal Society A. 2016;374(2065):20150202.

[5] Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al.. DINOv2: Learning Robust Visual Features Without Supervision; 2023. Available from: https://arxiv.org/abs/2304.07193.

[6] Zilberstein N, Mardani M, Segarra S. Repulsive Latent Score Distillation for Solving Inverse Problems; 2024. Available from: https://arxiv.org/abs/2406.16683.

[7] Poole B, Jain A, Barron JT, Mildenhall B. DreamFusion: Text-to-3D using 2D Diffusion; 2022. Available from: https://arxiv.org/abs/2209.14988.

[8] Wang Z, Lu C, Wang Y, Bao F, Li C, Su H, et al.. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation; 2023. Available from: https://arxiv.org/abs/2305.16213.

[9] Lin CH, Gao J, Tang L, Takikawa T, Zeng X, Huang X, et al.. Magic3D: High-Resolution Text-to-3D Content Creation; 2023. Available from: https://arxiv.org/abs/2211.10440.

[10] Hong Y, Zhang K, Gu J, Bi S, Zhou Y, Liu D, et al. LRM: Large Reconstruction Model for Single Image to 3D. In: International Conference on Learning Representations (ICLR); 2024. Available from: https://openreview.net/forum?id=sllU8vvsFF.

[11] Ho J, Jain A, Abbeel P. Denoising Diffusion Probabilistic Models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33; 2020. p. 6840-51.

[12] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Nets. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 27; 2014. Available from: https://papers.nips.cc/paper/5423-generative-adversarial-nets.

[13] Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. Communications of the ACM. 2021;65(1):99-106.

[14] Mescheder L, Oechsle M, Niemeyer M, Nowozin S, Geiger A. Occupancy Networks: Learning 3D Reconstruction in Function Space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019. p. 4460-70. Available from: https://arxiv.org/abs/1812.03828.

[15] Park JJ, Florence P, Straub J, Newcombe R, Lovegrove S. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019. p. 165-74. Available from: https://arxiv.org/abs/1901.05103.

[16] Seo J, Jang W, Kwak MS, Kim H, Ko J, Kim J, et al. Let 2D Diffusion Model Know 3D-Consistency for Robust Text-to-3D Generation. In: International Conference on Learning Representations (ICLR); 2024. Available from: https://openreview.net/forum?id=UbxWjq0UO2.

[17] Xie D, Li J, Tan H, Sun X, Shu Z, Zhou Y, et al.. Carve3D: Improving Multi-view Reconstruction Consistency for Diffusion Models with RL Finetuning; 2024. Available from: https://arxiv.org/abs/2312.13980.

[18] Liu Q, Wang D. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 29; 2016. Available from: https://papers.nips.cc/paper/6338-stein-variational-gradient-descent-a-general-purpose-bayesian-inference-algorithm.

[19] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-Resolution Image Synthesis with Latent Diffusion Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. p. 10684-95.

[20] Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35; 2022. p. 36589-602.

[21] Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical Text-Conditional Image Generation with CLIP Latents; 2022. Available from: https://arxiv.org/abs/2204.06125.

[22] Wang H, Du X, Li J, Yeh RA, Shakhnarovich G. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation; 2022. Available from: https://arxiv.org/abs/2212.00774.

[23] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning Transferable Visual Models from Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML); 2021. p. 8748-63.

[24] Borji A. Pros and Cons of GAN Evaluation Measures: New Developments. Computer Vision and Image Understanding. 2022;215:103329. Available from: https://doi.org/10.1016/j.cviu.2021.103329.

[25] Stein G, Cresswell JC, Hosseinzadeh R, Sui Y, Ross BL, Villecroze V, et al. Exposing Flaws of Generative Model Evaluation Metrics and Their Unfair Treatment of Diffusion Models. In: Advances in Neural Information Processing Systems (NeurIPS); 2023. Available from: https://openreview.net/forum?id=08zf7kTOoh.

[26] Jiang C, Zeng Y, Hu T, Xu S, Zhang W, Xu H, et al.. JointDreamer: Ensuring Geometry Consistency and Text Congruence in Text-to-3D Generation via Joint Score Distillation; 2024. Available from: https://arxiv.org/abs/2407.12291.

[27] Chen R, Chen Y, Jiao N, Jia K. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation; 2023. Available from: https://arxiv.org/abs/2303.13873.

[28] Liu M, Shi R, Chen L, Zhang Z, Xu C, Wei X, et al.. One-2-3-45++: Fast Single Image to 3D Objects with Consistent Multi-View Generation and 3D Diffusion; 2023. Available from: https://arxiv.org/abs/2311.07885.

[29] Shi R, Chen H, Zhang Z, Liu M, Xu C, Wei X, et al.. Zero123++: A Single Image to Consistent Multi-view Diffusion Base Model; 2023. Available from: https://arxiv.org/abs/2310.15110.

[30] Chen Z, Wang F, Wang Y, Liu H. Text-to-3D using Gaussian Splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024. p. 21535-45. Available from: https://openaccess.thecvf.com/content/CVPR2024/papers/Chen_Text-to-3D_using_Gaussian_Splatting_CVPR_2024_paper.pdf.

[31] Wang Q, Qi M, Huang Z, Zhang Y, Li Y, Wang Y, et al. Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2023. p. 19999-20009. Available from: https://arxiv.org/abs/2303.12789.

[32] Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, et al.. Emerging Properties in Self-Supervised Vision Transformers; 2021. Available from: https://arxiv.org/abs/2104.14294.

[33] Ilharco G, Wortsman M, Gontijo-Lopes R, et al.. OpenCLIP: An Open-Source Implementation of CLIP; 2021. https://github.com/mlfoundations/open_clip.

[34] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018. p. 586-95. Available from: https://arxiv.org/abs/1801.03924.

[35] Lakens D. Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-tests and ANOVAs. Frontiers in Psychology. 2013;4:863.

[36] Lance CE, Vandenberg RJ. Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences. Taylor & Francis; 2009.

[37] Deb K. Multi-Objective Optimization Using Evolutionary Algorithms. Wiley; 2001.

[38] Miettinen K. Nonlinear Multiobjective Optimization. Kluwer Academic Publishers; 1999.

[39] Deb K, Pratap A, Agarwal S, Meyarivan T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation. 2002;6(2):182-97.

[40] Wu T, Yang G, Li Z, Zhang K, Liu Z, Guibas L, et al. GPT-4V(ision) is a Human-Aligned Evaluator for Text-to-3D Generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024. p. 22227-38.

[41] Ye J, Liu F, Li Q, Wang Z, Wang Y, Wang X, et al. DreamReward: Text-to-3D Generation with Human Preference. In: European Conference on Computer Vision (ECCV). Springer; 2024. p. 259-76.

Declarations

Use of Generative AI. This dissertation made limited use of generative AI tools (OpenAI GPT-5) for drafting and refining technical explanations, generating LaTeX code, and improving clarity of writing. All technical content, methodology, and results were implemented and validated by the author. Generative AI was not used for producing experimental results, figures, or evaluation metrics. AI-assisted coding tools (GitHub Copilot, Cursor Agent) are acknowledged in the project repository (`CONTRIBUTORS.md`, `README.md`).

Ethical Considerations. The project builds on pretrained models (Stable Diffusion, DINOv2, OpenCLIP) licensed for research use. These models may encode social biases; care was taken to use prompts responsibly and avoid sensitive content. No human subjects or sensitive data were involved, and no ethical approval was required. All work was conducted under research-only licenses, with no commercial deployment intended.

Sustainability. Training was conducted on an NVIDIA A100 80 GB GPU. To reduce environmental and computational cost, mixed precision and gradient checkpointing were used, hyperparameter sweeps were limited to essential ranges, and renderings for debugging were stored at 256×256 resolution. Models and checkpoints were reused to avoid redundant computation, and evaluation metrics were logged systematically to enable post-hoc analysis.

The dissertation research spanned from late April to mid-September 2025, with intensive GPU usage concentrated in roughly ten days of development and experimentation. In total, about 240 GPU hours were consumed on a single A100, corresponding to $\approx$96 kWh of electricity. Using the UK grid average (0.233 $kgCO_2$/kWh), this equates to an estimated carbon footprint of 22.4 $kgCO_2e$ — roughly comparable to a 95 km car journey (e.g. London to Cambridge) or a one-way domestic flight from London to Manchester.

Availability of Data and Code. The full training and evaluation codebase, configuration files, and selected generated outputs will be made available on GitHub. Large pretrained weights (e.g. Stable Diffusion, DINOv2) are not included but can be obtained from their official repositories. Access to the full repository has been granted to the supervisor and second marker. All reproducibility artefacts accompanying this dissertation are available at: https://github.com/sijeong-kim/3D-Generation.

Appendix A

Additional Methodology

A.1 Compute Environment

Table A.1 and Table A.2 report the hardware and software configurations of the training host. These settings underpin all experiments reported in Chapter 4. GPU memory and timing logs were collected automatically during training; only aggregate results are reported in the main text.

Table A.1: Hardware configuration of the training host.

Component	Specification
CPU	AMD Ryzen Threadripper PRO 5955WX (16 cores / 32 threads)
RAM	512 GiB DDR4
GPU	NVIDIA A100 80 GB (PCIe)
Storage	NVMe SSD (scratch) + HDD (archive)

Table A.2: Software stack used in all experiments.

Component	Version / Build
OS	Ubuntu 22.04 LTS (64-bit)
Python	3.10.x
CUDA toolkit	12.8
NVIDIA driver	560.xx
cuDNN	9.1.0.2
PyTorch	2.8.0+cu128
OpenCLIP	ViT-bigG-14 (Ilharco et al., 2021)
DINOv2	`facebook/dinov2-base`, last-layer embeddings (Oquab et al., 2023)
Compiler toolchain	GCC 11.x / CMake 3.22+

Mixed precision and memory. Unless otherwise stated, training used automatic mixed precision (AMP, bfloat16 where supported) with gradient scaling enabled. Peak memory was recorded via `torch.cuda.max_memory_allocated()` and cross-checked with `nvidia-smi` sampling.

A.2 Determinism and Reproducibility

This section expands on reproducibility safeguards beyond what is reported in the methodology (Chapter 3). The following items ensure that results in Chapter 4 can be replicated exactly:

- **Seeds.** Global seeds were set for Python/NumPy/PyTorch; per-particle initial seeds followed `seed + 10^6 * j` (see Section 4.2.5).
- **Determinism.** `torch.backends.cudnn.deterministic=True` and `benchmark=False` were used for deterministic convs; operations known to be non-deterministic on the given CUDA/cuDNN stack are noted in the code repository README.

- **Precision.** AMP used `GradScaler` with default hysteresis; kernels excluded from AMP are whitelisted in config.
- **Environment capture.** For each run we store: git commit hash, `pip freeze`, CUDA/cuDNN versions, GPU name/VRAM, and a 64-bit config hash of the YAML file.
- **Artifacts.** YAML config, training/eval logs, checkpoints, and plots are archived per-run; filenames include the config hash for cross-checking.

A.3 Terminology and Label Mapping

To aid reproducibility, Table A.3 aligns implementation keys with the terminology and labels used in the dissertation. This mapping is particularly relevant when interpreting legacy figure exports (e.g. Figures 4.4 and 4.5), where `COS` should be read as equivalent to the *Shifted-cosine* kernel described in Sec. 3.4.

Table A.3: Code–paper terminology mapping.

Code key/value	Paper term	Figure label
`repulsion_type=rlsd`	RLSD (repulsion mechanism)	RLSD
`repulsion_type=svgd`	SVGD	SVGD
`kernel_type=rbf`	RBF kernel	RBF
`kernel_type=cosine`	cosine kernel (a.k.a. shifted–cosine)	cosine / COS
`guidance_scale` (*SD*)	Classifier-free guidance (CFG scale)	CFG
`lambda`	repulsion strength λ	λ
`beta`	kernel temperature β	β
`feature_layer=last`	last-layer DINOv2 features	last (DINOv2)
`clip_backend=openclip`	OpenCLIP (default)	OpenCLIP
`clip_backend=hf`	HuggingFace CLIP (alt.)	HF CLIP
combination notation	RLSD–RBF, RLSD–cosine, etc.	RLSD–RBF, RLSD–COS

Some legacy figures use the shorthand `COS`; this corresponds to the *cosine* kernel described in Sec. 3.4.

Notation. We use en-dashes for method–kernel combinations (e.g. RLSD–RBF). Greek symbols are rendered as λ (repulsion strength) and β (kernel temperature). All kernel names in the text are normalised to *cosine* or *shifted–cosine*; the label `COS` should be read as equivalent to *cosine*.

Appendix B

Additional Evaluation

This appendix provides extended experimental details supporting Chapter 4. It includes default configurations, ablation studies, prompt-wise comparisons, multi-view visualisations, efficiency statistics, and full documentation of the human study protocol. Throughout, we report three main metrics: $\mathcal{F}$ (fidelity), $\mathcal{D}$ (diversity), and $\mathcal{C}$ (cross-view consistency).

B.1 Experiment Settings and Parameters

The tables below expand on Chapter 4, providing full YAML-style defaults and the final configuration used in Exp6.

Table B.1: Default settings used throughout unless ablated in a named experiment. Note: `guidance_scale` is the Stable Diffusion guidance scale in code (SD), but reported in plots and text as the classifier-free guidance (CFG) scale.

Category	**Default value (unless stated)**
Training iterations	`iters=1000`; `schedule_iters=1500`
Guidance (SD; reported as CFG scale)	`guidance_scale=50` (ablated in Exp4)
Repulsion (ours)	`repulsion_type=rlsd`, `kernel_type=rbf`
Repulsion strength	$\lambda = 1000$ (Exp2–3 sweep)
Kernel temperature	`rbf_beta=1.0` (Exp5 sweep; final uses 0.5)
Feature space	DINOv2 (`facebook/dinov2-base`), `feature_layer=last`
Particles per prompt	`num_particles=8`
Diffusion controls	`force_same_t=True`, `force_same_noise=True`
Timestep policy	`anneal_timestep=True`
Background (train)	`invert_bg_prob=0.5`
Evaluation render	`eval_H=512`, `eval_W=512`, `num_views=8`
Camera radius (eval)	`bulldozer`: 4.5; `hamburger`: 3.0; `icecream/cactus/tulip/sundae`: 4.0
Densification	`density_start_iter=0`, `density_end_iter=3000`, `densification_interval=50`, `densify_grad_threshold=0.01`
Opacity	`opacity_reset_interval=700`, `opacity_lr=0.01`
Seeds	{42, 123, 456, 789} (Exp2 uses 42 only)
Metrics/logging	losses/efficiency every 10 iters; quantitative metrics every 50 iters; LPIPS off

Table B.2: Configuration of the best-performing model (`exp6_ours_best`) used in Exp6.

Setting	**Value**
Repulsion	`repulsion_type=rlsd`, `kernel_type=rbf`
Strength & temperature	$\lambda = 1000$, `rbf_beta=0.5`
Feature space	DINOv2 (`facebook/dinov2-base`), `feature_layer=last`
Guidance (SD)	`guidance_scale=50`
Training schedule	`iters=1000`; `schedule_iters=1500`; `anneal_timestep=True`
Diffusion controls	`force_same_t=True`, `force_same_noise=True`
Particles per prompt	`num_particles=8`
Evaluation render	`eval_H=512`, `eval_W=512`, `num_views=8`
Camera radius (eval)	`bulldozer`: 4.5; `hamburger`: 3.0; `icecream/cactus/tulip/sundae`: 4.0
Densification	`density_start_iter=0`, `density_end_iter=3000`, `densification_interval=50`, `densify_grad_threshold=0.01`
Opacity	`opacity_reset_interval=700`, `opacity_lr=0.01`
Seeds	{42, 123, 456, 789}
Prompts (Exp6)	`bulldozer`, `icecream`, `cactus`, `tulip`, `hamburger`, `sundae`

B.2 Baseline

Baseline results correspond to Section 4.3, comparing multi-particle training without repulsion (*wo*; shared t, ϵ) against the independent-noise baseline. Values are reported as mean ± SE over seeds.

Table B.3: Baseline settings: multi-particle (wo) vs. independent-noise. Values are mean ± SE over seeds. The last row reports averages across prompts.

(a) Baseline (wo; shared t, ϵ)

Prompt	$\mathcal{F}\uparrow$	$\mathcal{D}\uparrow$	$\mathcal{C}\uparrow$
bulldozer	0.401 ± 0.003	0.190 ± 0.003	0.816 ± 0.003
tulip	0.394 ± 0.002	0.100 ± 0.005	0.895 ± 0.004
icecream	0.419 ± 0.001	0.151 ± 0.022	0.811 ± 0.004
cactus	0.379 ± 0.004	0.111 ± 0.007	0.892 ± 0.004
Overall	0.398 ± 0.002	0.138 ± 0.009	0.853 ± 0.004

(b) Independent-noise baseline

Prompt	$\mathcal{F}\uparrow$	$\mathcal{D}\uparrow$	$\mathcal{C}\uparrow$
bulldozer	0.400 ± 0.003	0.209 ± 0.011	0.811 ± 0.009
tulip	0.395 ± 0.001	0.119 ± 0.004	0.893 ± 0.004
icecream	0.419 ± 0.002	0.202 ± 0.012	0.815 ± 0.004
cactus	0.378 ± 0.003	0.131 ± 0.007	0.891 ± 0.002
Overall	0.398 ± 0.002	0.165 ± 0.009	0.853 ± 0.005

B.3 Additional Ablation Studies

This section expands on Chapter 4, evaluating the impact of feature-layer choice, particle count N, initial Gaussian count, and opacity learning rate. Results are averaged over prompts (seed=42 for efficiency).

Table B.4: Additional ablations (seed=42). Each subtable reports averages across prompts; higher is better for $\mathcal{F}, \mathcal{D}, \mathcal{C}$.

(a) Feature-layer choice (DINOv2)

Layer	$\mathcal{F}$	$\mathcal{D}$	$\mathcal{C}$
early	0.374	0.332	0.861
mid	0.390	0.258	0.849
last	0.389	0.271	0.833

(b) Number of particles N

N	$\mathcal{F}$	$\mathcal{D}$	$\mathcal{C}$
2	0.394	0.233	0.838
4	0.396	0.261	0.844
8	0.386	0.265	0.833

(c) Number of points (initial)

Points	$\mathcal{F}$	$\mathcal{D}$	$\mathcal{C}$
1000	0.397	0.281	0.842
3000	0.399	0.271	0.834
5000	0.398	0.287	0.834

(d) Opacity learning rate

LR	$\mathcal{F}$	$\mathcal{D}$	$\mathcal{C}$
0.005	0.398	0.285	0.839
0.010	0.398	0.272	0.845
0.050	0.389	0.249	0.848

Table B.5: Consolidated results across Experiments 1–6. Values are mean ± SE over prompts × seeds (N=8, V=8). Δ columns denote absolute differences vs. the multi-particle (wo) baseline.

Exp	Method	$\mathcal{F}\uparrow$	$\Delta\mathcal{F}$	$\mathcal{D}\uparrow$	$\Delta\mathcal{D}$	$\mathcal{C}\uparrow$	$\Delta\mathcal{C}$
–	Baseline (wo)	0.398 ± 0.001	–	0.138 ± 0.002	–	0.853 ± 0.001	–
1	SVGD–COS	0.369 ± 0.001	-0.030	0.181 ± 0.002	+0.043	0.846 ± 0.002	-0.007
1	SVGD–RBF	0.380 ± 0.000	-0.018	0.254 ± 0.002	+0.116	0.835 ± 0.001	-0.019
1	RLSD–COS	0.393 ± 0.000	-0.005	0.198 ± 0.002	+0.060	0.845 ± 0.002	-0.008
1	RLSD–RBF	0.388 ± 0.000	-0.011	0.267 ± 0.002	+0.129	0.831 ± 0.001	-0.022
2	$\lambda = 1$	0.398	-0.001	0.136	-0.002	0.859	+0.004
2	$\lambda = 10$	0.398	-0.002	0.143	+0.005	0.855	0.000
2	$\lambda = 100$	0.396	-0.003	0.195	+0.057	0.848	-0.007
2	$\lambda = 1000$	0.387	-0.012	0.272	+0.134	0.828	-0.027
2	$\lambda = 10000$	0.290	-0.109	0.316	+0.178	0.860	+0.005
3	$\lambda = 600$	0.390 ± 0.001	-0.008	0.250 ± 0.002	+0.112	0.835 ± 0.001	-0.018
3	$\lambda = 800$	0.389 ± 0.001	-0.009	0.259 ± 0.002	+0.121	0.831 ± 0.001	-0.022
3	$\lambda = 1000$	0.388 ± 0.001	-0.010	0.267 ± 0.002	+0.129	0.832 ± 0.001	-0.021
3	$\lambda = 1200$	0.386 ± 0.001	-0.012	0.273 ± 0.004	+0.136	0.833 ± 0.002	-0.020
3	$\lambda = 1400$	0.384 ± 0.001	-0.014	0.286 ± 0.004	+0.148	0.828 ± 0.002	-0.025
4	Guidance=30	0.384 ± 0.001	-0.015	0.299 ± 0.002	+0.162	0.831 ± 0.001	-0.022
4	Guidance=50	0.387 ± 0.001	-0.011	0.259 ± 0.002	+0.121	0.831 ± 0.001	-0.023
4	Guidance=70	0.388 ± 0.001	-0.011	0.255 ± 0.002	+0.118	0.831 ± 0.001	-0.022
4	Guidance=100	0.386 ± 0.001	-0.012	0.243 ± 0.002	+0.105	0.834 ± 0.002	-0.019
5	$\beta_{\text{RBF}} = 0.5$	0.388 ± 0.001	-0.010	0.262 ± 0.002	+0.124	0.831 ± 0.002	-0.022
5	$\beta_{\text{RBF}} = 1.0$	0.388 ± 0.001	-0.011	0.263 ± 0.003	+0.125	0.833 ± 0.002	-0.020
5	$\beta_{\text{RBF}} = 1.5$	0.387 ± 0.001	-0.012	0.260 ± 0.002	+0.122	0.833 ± 0.001	-0.020
5	$\beta_{\text{RBF}} = 2.0$	0.392 ± 0.001	-0.006	0.243 ± 0.003	+0.105	0.835 ± 0.002	-0.019
6	Final (Ours)	0.391 ± 0.004	-0.006	**0.262 ± 0.006**	+0.130	0.831 ± 0.007	-0.025

B.4 Consolidated Results Across Experiments

Table B.5 aggregates Exps 1–6, with absolute differences (Δ) computed against the multi-particle baseline. Bold values denote the best-performing configuration (Exp6, RLSD–RBF).

B.5 Prompt-wise Comparison

Table B.6 reports per-prompt means ± SE across seeds, complementing the main prompt-wise comparison.

Table B.6: Prompt-wise comparison of final setting (Ours) vs. baseline. Values are mean ± SE over seeds.

Prompt	**Baseline**			**Ours**		
	$\mathcal{F}$	$\mathcal{D}$	$\mathcal{C}$	$\mathcal{F}$	$\mathcal{D}$	$\mathcal{C}$
bulldozer	0.4006 ± 0.0015	0.1891 ± 0.0025	0.8146 ± 0.0033	0.3790 ± 0.0025	0.2578 ± 0.0042	0.8021 ± 0.0047
cactus	0.3770 ± 0.0017	0.1093 ± 0.0035	0.8946 ± 0.0029	0.3765 ± 0.0008	0.2283 ± 0.0017	0.8773 ± 0.0047
hamburger	0.4139 ± 0.0015	0.0942 ± 0.0021	0.9037 ± 0.0007	0.4102 ± 0.0024	0.2565 ± 0.0084	0.8592 ± 0.0009
icecream	0.4194 ± 0.0003	0.1433 ± 0.0099	0.8115 ± 0.0025	0.4180 ± 0.0015	0.2881 ± 0.0080	0.8047 ± 0.0033
sundae	0.3794 ± 0.0011	0.1507 ± 0.0065	0.8228 ± 0.0044	0.3775 ± 0.0004	0.3080 ± 0.0077	0.7931 ± 0.0020
tulip	0.3931 ± 0.0007	0.1064 ± 0.0043	0.8905 ± 0.0021	0.3847 ± 0.0007	0.2327 ± 0.0037	0.8508 ± 0.0023

B.6 Multi-view Visualisations

The visualisations in Figures B.1 to B.6 show qualitative results for the six prompts listed in Table 4.1. Each figure displays N=8 particles and V=8 uniformly distributed camera views.

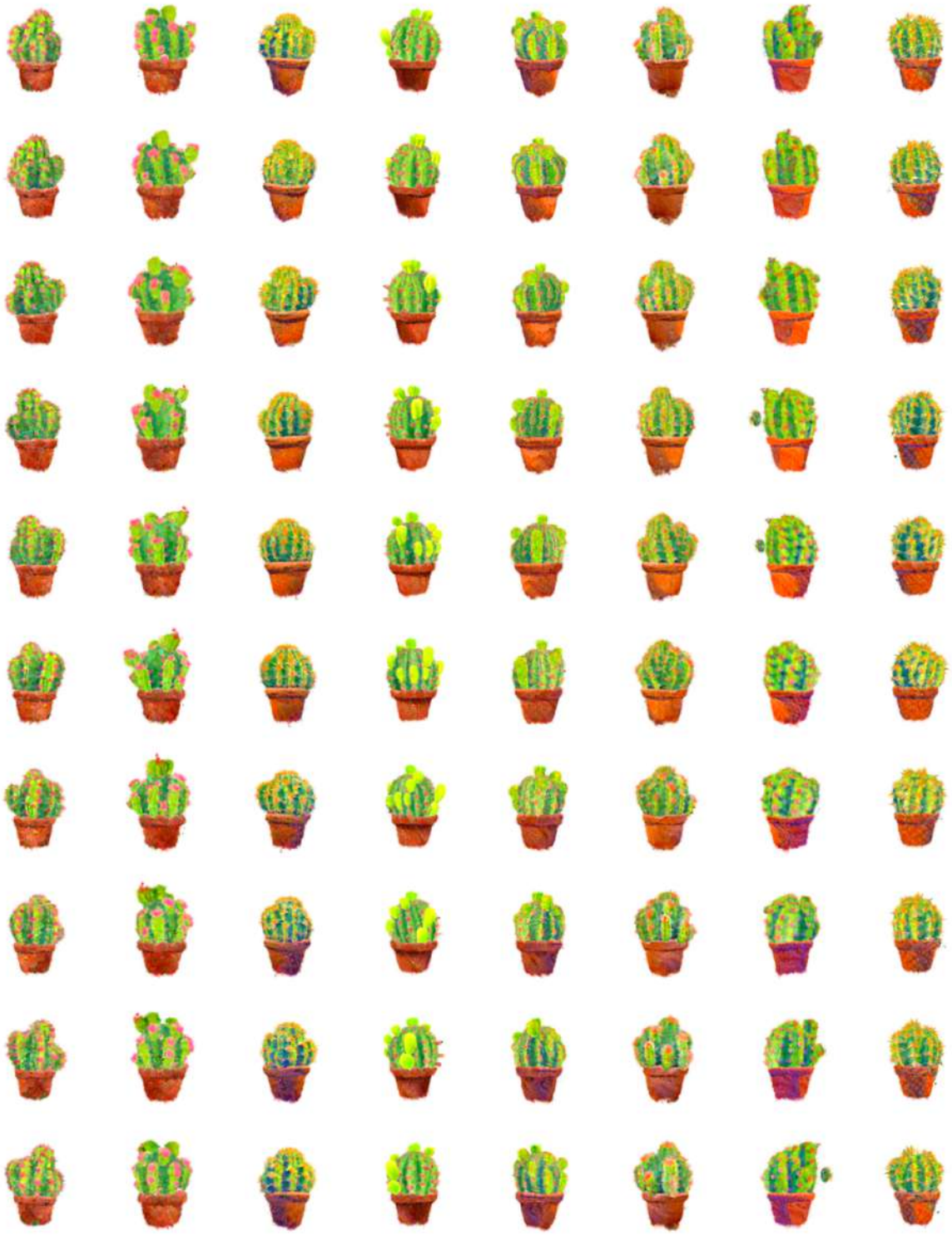


Figure B.1: Multi-view results for the best-performing model (`exp6_ours_best`) on the `cactus` prompt (seed=42). Each row corresponds to a novel camera view (uniformly distributed azimuths at 0° elevation), and each column corresponds to a different particle instance. This layout illustrates both cross-particle diversity (rows) and cross-view consistency (columns).

Figure B.2: Multi-view results for the best-performing model (`exp6_ours_best`) on the `hamburger` prompt (seed=42). Layout as in Figure B.1.

Figure B.3: Multi-view results for the best-performing model (`exp6_ours_best`) on the `icecream` prompt (seed=42). Layout as in Figure B.1.

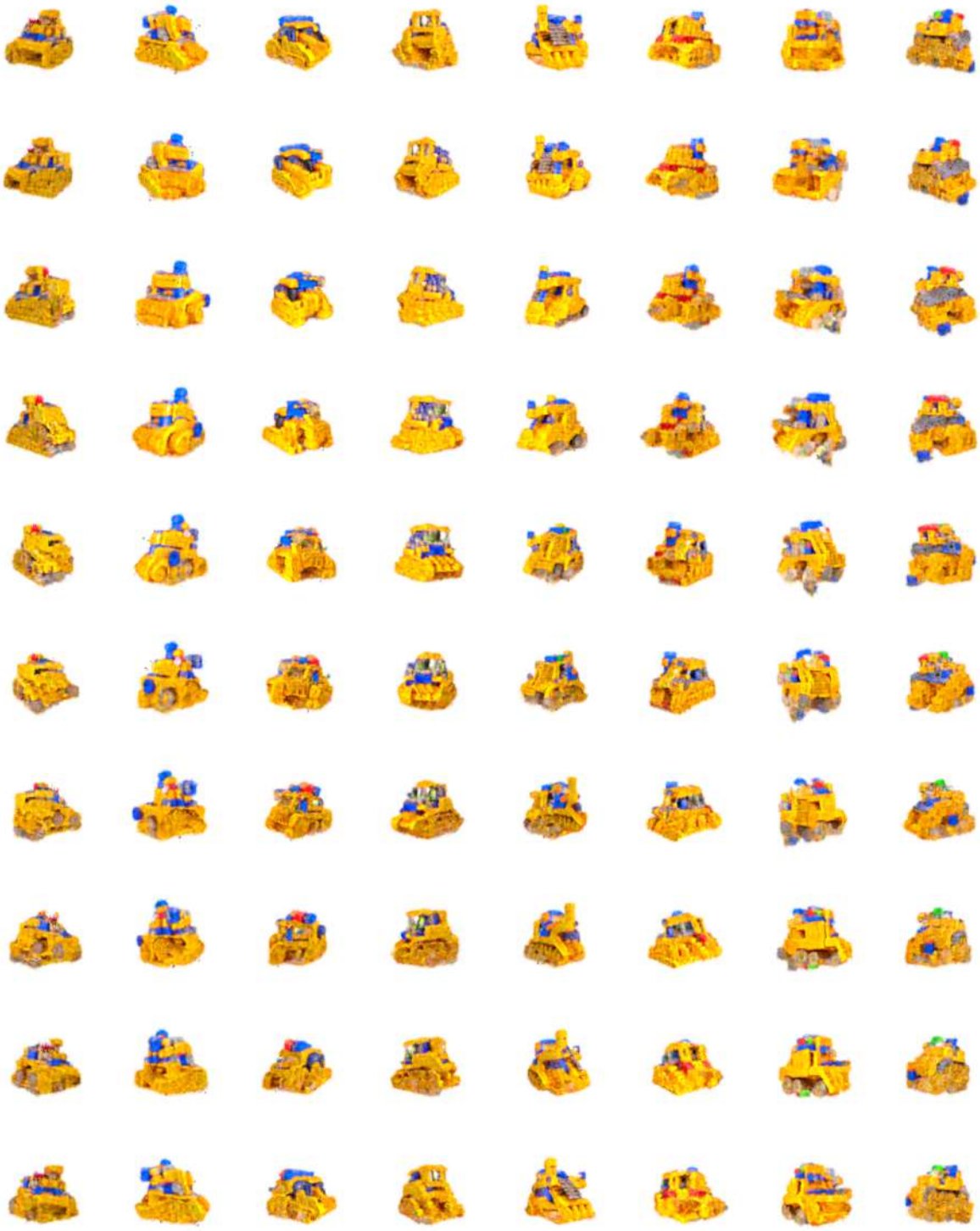


Figure B.4: Multi-view results for the best-performing model (`exp6_ours_best`) on the `bulldozer` prompt (seed=42). Layout as in Figure B.1.

Figure B.5: Multi-view results for the best-performing model (`exp6_ours_best`) on the `sundae` prompt (seed=42). Layout as in Figure B.1.

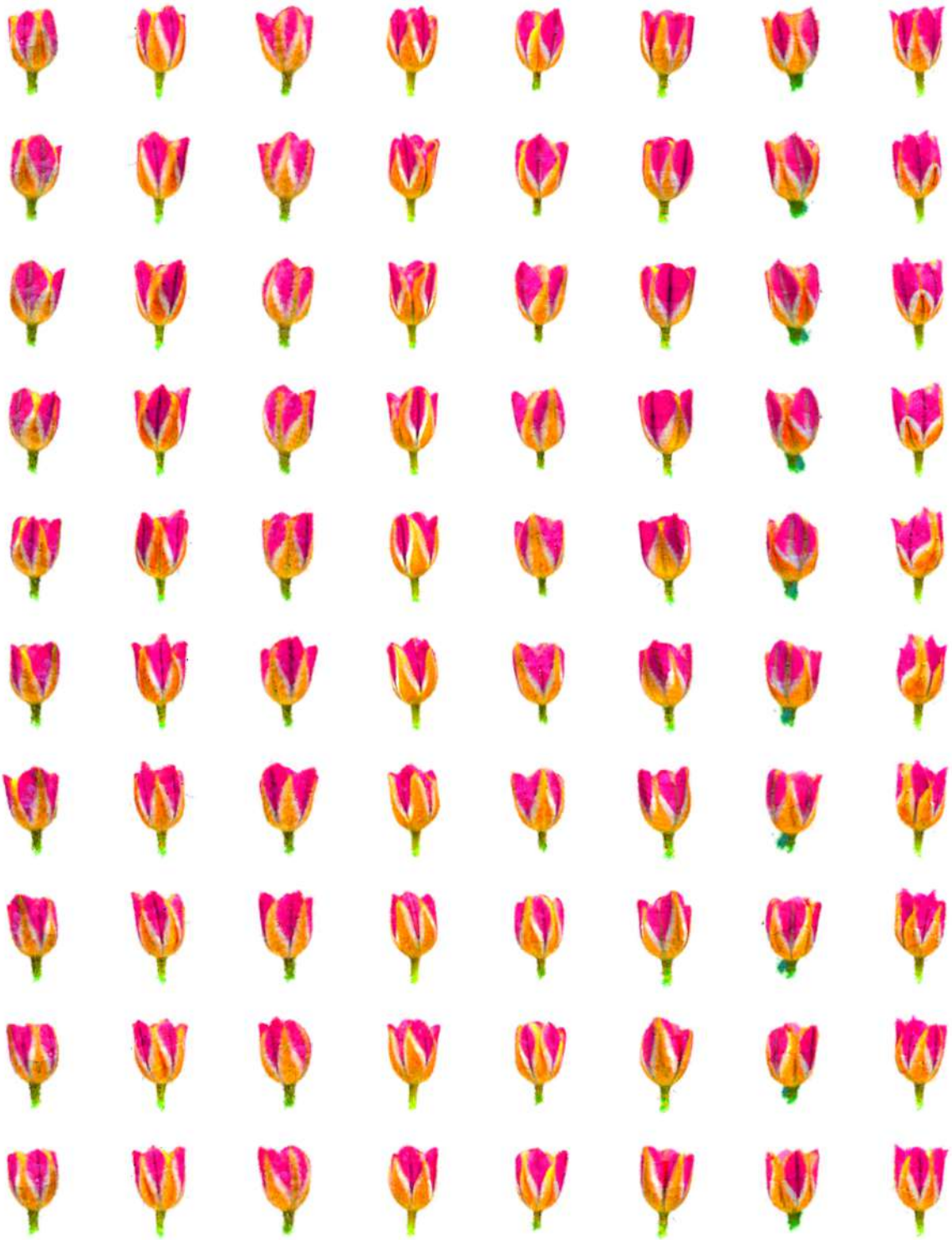
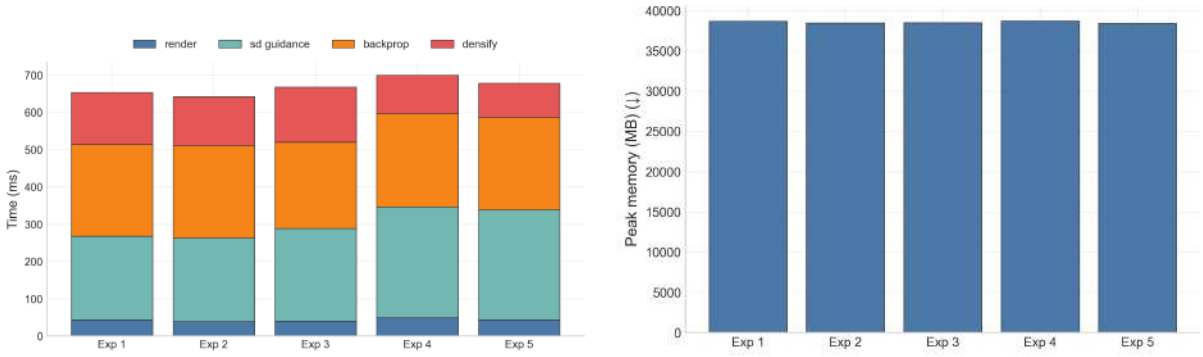


Figure B.6: Multi-view results for the best-performing model (`exp6_ours_best`) on the `tulip` prompt (seed=42). Layout as in Figure B.1.



(a) Per-step runtime breakdown

(b) Peak memory usage

Figure B.7: **Efficiency across experiments.** SD-guidance and backprop dominate runtime; wall time remains ∼640–700 ms/step. Peak memory is stable at ∼38–39 GB.

Table B.7: Exp6: Efficiency statistics (mean ± std; baseline = multi-particle (wo), N=8).

Metric	Baseline	Ours	Δ
Peak memory (MB)	37006 ± 45	38382 ± 26	+1376
Allocated memory (MB)	23066 ± 19	23236 ± 11	+170
Throughput (px/s)	976,831 ± 10,070	867,183 ± 3,093	-11%
Step time (ms)	646 ± 39	634 ± 2	-12

B.7 Efficiency and Resource Use

Efficiency analysis complements Figure 4.21. Figure B.7 shows per-step runtime and peak memory usage across experiments. Table B.7 reports detailed statistics for Exp6.

B.8 Human Study Protocol

Details expand on Section 4.5.4. We followed standard practice [40, 41], using pairwise realism preferences and 5-point Likert diversity ratings. Prompt order and A/B assignment were randomised per participant. A total of $N = 41$ participants completed the study.

B.9 Statistical Analysis of Human Study

Paired t-tests were conducted on Likert differences (ours – baseline), confirming significantly higher diversity ratings for our method (all $p < 0.001$, Bonferroni corrected). Effect sizes (Cohen's d) were consistently large ($d > 0.8$).

Realism preferences were analysed separately as categorical outcomes (baseline / ours / tie), for which no systematic bias was observed.

To test whether the observed diversity differences were statistically significant, we conducted paired t-tests across participants. For each prompt, every participant provided two Likert ratings (baseline vs. ours). We subtracted the baseline rating from the corresponding rating for our method, yielding a paired difference score per participant. The null hypothesis was that the mean difference equals zero. This approach follows standard practice for analysing repeated-measures Likert data [35, 36, 40, 41].

Across prompts, the paired t-tests confirmed significantly higher diversity ratings for our method (all $p < 0.001$ after Bonferroni correction). Realism preferences were analysed separately as categorical outcomes (baseline / ours / same), for which no systematic bias was observed.

B.10 Ethics, Consent, and Data Protection

This expands on the Declarations (Section 5.3), detailing consent, anonymity, and data handling. The study was classified as low-risk by the Departmental ethics lead.

This user study was classified as low-risk and conducted in accordance with the Departmental guidelines for human participant research. Participation was voluntary, restricted to individuals aged 18 or above, and proceeded only after participants explicitly selected "I consent" on an informed-consent screen. Participants could withdraw at any time by closing the survey; no identifying or sensitive personal data were collected.

Anonymity and confidentiality. No names, email addresses, IP addresses, or device identifiers were collected. All responses were stored in anonymised form and analysed only in aggregate.

Data handling and retention. Survey responses were stored on institutionally managed storage, accessible only to the research team. Data will be retained for up to 12 months after publication for verification and will then be deleted. No data will be shared publicly beyond aggregated, anonymised statistics and figures reported in this thesis.

Risks, benefits, and compensation. The study posed minimal risk and involved viewing computer-generated images and short animations. No compensation was offered. Participants could skip any item or withdraw without penalty.

B.11 Informed Consent Text (Survey)

Title: Evaluation of Visual Diversity in Text-to-3D Results

Purpose. You are invited to take part in a brief online study that compares two sets of 3D results generated by different algorithms. Your responses will help us evaluate perceived realism and diversity.

Procedures. You will be shown pairs of multi-view grids and short animations (A/B). For each pair, you will (i) choose which set appears more realistic and (ii) rate the diversity of each set on a 1–5 scale. The study takes approximately 5–10 minutes.

Eligibility. You must be 18 years of age or older.

Voluntary participation and withdrawal. Your participation is voluntary. You may discontinue at any time by closing the survey. You may skip any question you prefer not to answer.

Data and anonymity. We do not collect names, email addresses, IP addresses, or other identifiers. Responses are stored and analysed only in aggregate, anonymised form for research purposes.

Contact. If you have questions about this study, please contact `<sk2324@ic.ac.uk>`.

Consent. By selecting "I consent" below, you confirm that you are 18 or over, you have read and understood this information, and you voluntarily agree to participate. Selecting "I do not consent" will close the survey.