

Extending Data Priors Across Domains with Diffusion Distillation

David McAllister
Angjoo Kanazawa, Ed.
Alexei (Alyosha) Efros, Ed.

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2024-177

<http://www2.eecs.berkeley.edu/Pubs/TechRpts/2024/EECS-2024-177.html>

August 9, 2024



Copyright © 2024, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Acknowledgement

We thank Matthew Tancik, Jiaming Song, Riley Peterlinz, Ayaan Haque, Ethan Weber, Konpat Preechakul, Amit Kohli and Ben Poole for their helpful feedback and discussion.

Extending Data Priors Across Domains with Diffusion Distillation

by

David McAllister

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Master of Science

in

Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Angjoo Kanazawa, Chair
Alexei Efros

Spring 2024

Extending Data Priors Across Domains with Diffusion Distillation

Copyright 2024
by
David McAllister

Extending Data Priors Across Domains with Diffusion Distillation

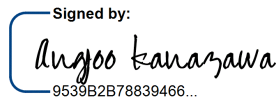
by David McAllister

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Signed by:

9539B2B78839466...

Professor Angjoo Kanazawa
Research Advisor

09/08/2024

(Date)

* * * * *

Signed by:

3999E536FB3D48F...

Professor Alexei (Alyosha) Efros
Second Reader

09/08/2024

(Date)

Abstract

Extending Data Priors Across Domains with Diffusion Distillation

by

David McAllister

Master of Science in Electrical Engineering and Computer Sciences

University of California, Berkeley

Angjoo Kanazawa, Chair

Score distillation sampling (SDS) has proven to be an important tool, enabling the use of large-scale diffusion priors for tasks operating in data-poor domains. Unfortunately, SDS has a number of characteristic artifacts that limit its usefulness in general-purpose applications. In this paper, we make progress toward understanding the behavior of SDS and its variants by viewing them as solving an optimal-cost transport path from a source distribution to a target distribution. Under this new interpretation, these methods seek to transport corrupted images (source) to the natural image distribution (target). We argue that current methods' characteristic artifacts are caused by (1) linear approximation of the optimal path and (2) poor estimates of the source distribution. We show that calibrating the text conditioning of the source distribution can produce high-quality generation and translation results with little extra overhead. Our method can be easily applied across many domains, matching or beating the performance of specialized methods. We demonstrate its utility in text-to-2D, text-based NeRF optimization, translating paintings to real images, optical illusion generation, and 3D sketch-to-real. We compare our method to existing approaches for score distillation sampling and show that it can produce high-frequency details with realistic colors.

Contents

Contents	i
List of Figures	ii
List of Tables	iv
1 Introduction	1
2 Related Work	3
3 Method	5
3.1 Experiments	11
4 Conclusion	16
4.1 Discussion	16
4.2 Closing Remarks	17
Bibliography	18
5 Appendix	23
5.1 Additional Experimental Setup	23
5.2 More Visual Results	23

List of Figures

3.1	Optimization with diffusion models as approximation of a Schrödinger Bridge Problem (SBP). (a) We propose to formulate optimization with diffusion models as bridging the distribution of the current optimized image x_θ to the target distribution under a dual-bridge framework (a). Current methods can be interpreted as approximating the optimal transport ϵ_{SBP}^* between these distributions via the difference between projections of a noised image $x_{\theta,t}$ onto the two distributions. This analysis reveals two sources of error: (1) these gradients are linear approximations of the optimal path, as illustrated in (a), and (2) the source distribution used for computing this approximation (<i>e.g.</i> , the unconditional distribution in SDS [46]) may not be aligned with the current distribution, illustrated in (b).	6
3.2	Comparison of SDS variants under our analysis. We illustrate the major gradient components of different SDS variants and provide a straightforward comparison with ϵ_{SBP}	8
3.3	Text-to-image generation results with COCO Captions. We compare different score distillation methods for generating images with COCO captions by optimizing a randomly initialized image. DDIM sampling indicates the lower bound that the diffusion model can achieve. VSD [68] and our method generate the least color artifacts while ours is more efficient than VSD.	11
3.4	Text-guided NeRF optimization with different score distillation methods. We make a fair comparison of SDS and VSD for text-to-3D generation. For each generation, we show three uniformly sampled views. SDS results like the cottage and pepper mill still suffer from over-saturation problems, while ours and VSD can produce realistic details, color, and texture.	13
3.5	Painting-to-Real comparison. We compare our gradient in optimization to image restoration and image-conditional generation baselines. While SDEdit produces convincing textures, its difficult to find a strength value that balances structure and quality. Other baselines fail to reproduce natural image quality, while our method produces the best combination of quality and faithfulness.	14
3.6	Painting-to-Real results. We show selected Painting-to-Real samples with diverse art styles and subjects. Initialization images are shown on the left, optimized images are shown on the right.	15

5.1	3D sketch-to-real. We introduce a conditional generation task in 3D where a coarse human-drawn mesh is optimized into a high-quality mesh. While SDS and our gradient both adhere to the prompt and shape conditions, our method produces higher fidelity colors and texture.	24
5.2	Diffusion illusions. We generate overlaid optic illusions with SDS and our method. While SDS suffers from color artifacts, our methods produce more details and proper color.	25
5.3	Ablation study of our method without stage 1. We show directly optimizing with y_{src} from the start could undermine the quality of the geometry and produce unnecessary content.	25
5.4	Additional comparison of text-guided NeRF optimization. We show more examples to compare with different distillation methods, SDS and VSD.	27

List of Tables

- 3.1 **Zero-shot FID comparison with different score distillation methods.** We report FID scores of text-to-image generation using 5K captions randomly sampled from the COCO dataset. The best score distillation result is indicated in **bold**, while the second best is underlined. 12
- 3.2 **Quantitative comparisons of NeRF optimization.** We measure the average CLIP similarity of rendered views using SDS, VSD and our experimental method. 13

Acknowledgments

We thank Matthew Tancik, Jiaming Song, Riley Peterlinz, Ayaan Haque, Ethan Weber, Konpat Preechakul, Amit Kohli and Ben Poole for their helpful feedback and discussion. This project is supported in part by a Google research scholar award and IARPA DOI/IBC No. 140D0423C0035. The views and conclusions contained herein are those of the authors and do not represent the official policies or endorsements of these institutions.

Chapter 1

Introduction

Diffusion models have shown tremendous success in modeling complex data distributions like images [49, 52, 3, 23], videos [57, 4] and robot action policies [13]. In domains where data is plentiful, they produce state-of-the-art results. Many data modalities, however, cannot enjoy the same scaling benefits due to their lack of sufficiently large datasets. In these cases, it is useful to exploit diffusion models trained on domains with rich data sources as a prior in an optimization framework. Score Distillation Sampling (SDS) [46, 67] and its variants [68, 20, 74] are a widely adopted way to optimize parametric images, *i.e.*, images produced by a model like NeRF, with a pre-trained diffusion model. Despite being applicable to a wide range of applications, SDS is also known to suffer from several significant artifacts, such as oversaturation and oversmoothing. As such, several variants have been proposed to alleviate these artifacts [68, 74, 32], often at the cost of efficiency, diversity, or other artifacts.

In this paper, we investigate the core issues with SDS by casting the class of score distillation optimization problems as a Schrödinger Bridge (SB) problem [53, 12, 11, 33], which finds the optimal transport between two distributions. The density flow formed by these mappings is transport-optimal, as defined in the SB problem. In an optimization framework, the difference between paired source and target samples, computed with an SB, can be used as a gradient to update the source. Su *et al*[63] have shown that this path can be explicitly solved using two pre-trained diffusion models. We show that one can also compose these models as an optimizer to approximate transport paths on the fly. Under this framework, we can understand SDS and its variants as approximating a source-to-target distribution bridge with the difference of two denoising directions. The denoising scores point to the source and target distributions respectively, with the source representing the current optimized image that updates with each optimization step.

This framing reveals two sources of errors. Our analysis reveals that SDS approximates the current distribution with the unconditional image distribution, which is not accurate and results in a *distribution mismatch error*. We show that recent SDS variants [68, 74, 32] can be seen as proposals to improve this distribution mismatch error.

Finally, our analysis motivates a simple method that rectifies the distribution mismatch issue without additional computational overhead. Our insight is that the large-scale text-to-image diffusion models learn from billions of caption-image pairs [54], where a breadth of image corruptions are present in their training sets. They are also equipped with powerful

pre-trained text encoders, which empower the models with zero-shot capacity in generating unseen concepts [51, 50]. As such, simply describing the current source distribution with text, even if it is not part of the real image manifold, can approximate the distribution of the current optimized image, leading to improved transport paths. Our simple and efficient solution can be easily applied to any existing application that uses SDS. We show that it consistently improves the visual quality in the desired domain. We comprehensively compare our approach with standard distillation sampling methods over several generation tasks, where our approach matches or outperforms the baselines.

Our contributions are as follows:

- We propose to cast the problem of using a pre-trained diffusion model as a prior in an optimization problem as solving the Schrödinger Bridge (SB) problem between two image distributions. Specifically, it can be seen as bridging the distribution of the current optimized image to the target distribution under a dual-bridge framework.
- We analyze recent SDS-based methods under the lens of our framework and explain the pros and cons of the individual methods.
- Our analysis motivates a simple yet effective alternative to SDS by using textual descriptions to specify the current optimized image distribution. It achieves consistently more realistic results than SDS, producing quality comparable with VSD [68] without its computational overhead. We compare various generation tasks to show its wall-clock efficiency and quality generations against state-of-the-art methods.

Chapter 2

Related Work

Score Distillation Sampling

Although modalities like 3D, 4D, sketch, and vector graphics (SVGs) lack the large-scale, diverse, and high-quality datasets needed to train a domain-specific diffusion model, previous works find it useful to exploit image or video as a proxy modality [26, 16]. By computing the gradient on a proxy representation with a pretrained model, optimization in the target modality is viable with differentiable mappings, e.g. differentiable rasterization [34] for SVGs or differentiable rendering [44] for 3D objects and scenes. The seminal method, Score Distillation Sampling (SDS) [46], first proposed to apply a pretrained text-to-image diffusion model for text-to-3D generation. However, it requires a high classifier-free guidance weight and, therefore, suffers from artifacts such as over-saturation and over-smoothing. Recent works have built upon SDS to adapt it for editing tasks [30, 20, 45, 29] or more broadly improve over the original SDS formulation [28, 1, 68, 75, 74, 76]. NFSD [28] and LMC-SDS [1] inspect the individual components of the SDS gradient and propose methods to rectify the high guidance weights. However, the over-saturation problem is mitigated but not fully resolved. VSD [68] formulates the problem as particle-based variational inference and proposes to train a LoRA [24] on the fly to estimate the score of proxy distribution. We presented a new framework that allows rethinking all the variants under the same lens. This framework also motivates a method that improves the quality of SDS without losing efficiency.

Visual Content Generation with SDS

Since SDS was developed for text-to-3D generation, it has also been adopted to generate various other visual content such as SVGs [18, 71], sketches [70], texture [43, 6, 7, 8, 73], typography [25], dynamic 4D scenes [2, 58, 38] and illusions [5]. Among these applications, text-to-3D has been the most active research direction. In addition to designing better distillation sampling methods [68, 75, 28], prior work has also studied the underlying 3D neural representations [72, 64, 36, 9] and leveraging multiview data to improve the 3D

consistency [55, 41, 40, 47, 76]. We note that these explorations are orthogonal to our study and should be able to work jointly with our method. In this paper, we looked into existing applications like text-based NeRF optimization, painting-to-real, and illusion generation. We also propose a new AR application called 3D sketch-to-real.

Chapter 3

Method

In this section, we present an analytical framework that casts the score distillation sampling (SDS) family of methods as instantiations of a Schrödinger Bridge problem. We show that many recent SDS based methods can be interpreted as an online solver for the problem. That is, each SDS optimization step is a first-order approximation of a dual diffusion bridge formed by two probability flow (PF) ODEs [63].

We analyze SDS and its variants under this general framework. Then, we present a simple solution based on the analysis, which leads to significant quality improvement with little extra computational overhead.

Background

Diffusion models define a forward “noising” process that degrades data samples \mathbf{x} gradually from the image distribution to noised samples \mathbf{z}_t , and eventually the i.i.d. Gaussian distribution [22, 60]. This process is indexed by timesteps t , where $t = 1$ indexes the full Gaussian noise distribution and $t = 0$ indexes the data distribution. A diffusion model, parameterized by ϕ , is then trained to reverse this encoding process, iteratively transforming the noise distribution into the data distribution with the score-matching objective:

$$\mathcal{L}_{\text{Diff}}(\phi, \mathbf{x}) = \mathbb{E}_{t \sim \mathcal{U}(0,1), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [w(t) \|\epsilon_\phi(\alpha_t \mathbf{x} + \sigma_t \epsilon; y, t) - \epsilon\|_2^2], \quad (3.1)$$

where y is a conditioning text prompt, and α_t and σ_t are hyperparameters from the predefined noise schedule.

Probability Flow ODE. Denoising score matching [62, 27, 59] shows that the diffusion model denoising prediction can be rewritten as a score vector field:

$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) = -\frac{1}{\sqrt{1 - \alpha_t}} \epsilon_t. \quad (3.2)$$

Because of its special connection to marginal probability densities, the resulting ODE is named the probability flow (PF) ODE with the following expression:

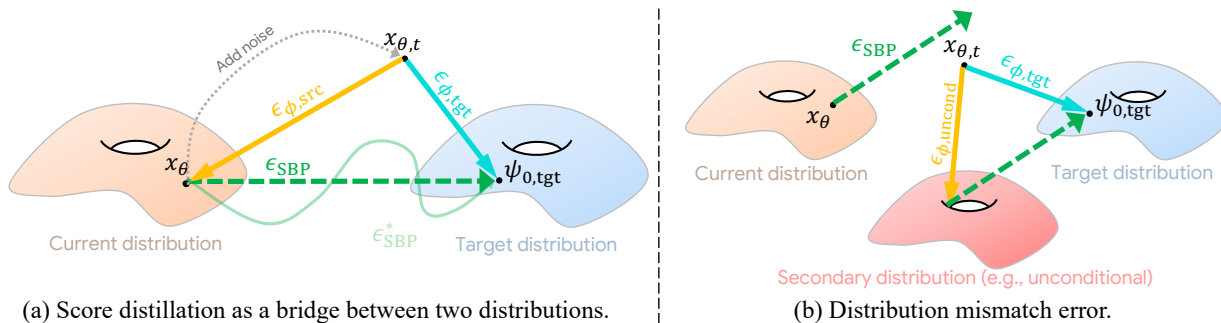


Figure 3.1: **Optimization with diffusion models as approximation of a Schrödinger Bridge Problem (SBP).** (a) We propose to formulate optimization with diffusion models as bridging the distribution of the current optimized image x_θ to the target distribution under a dual-bridge framework (a). Current methods can be interpreted as approximating the optimal transport ϵ_{SBP}^* between these distributions via the difference between projections of a noised image $x_{\theta,t}$ onto the two distributions. This analysis reveals two sources of error: (1) these gradients are linear approximations of the optimal path, as illustrated in (a), and (2) the source distribution used for computing this approximation (e.g., the unconditional distribution in SDS [46]) may not be aligned with the current distribution, illustrated in (b).

$$dx = [f(\mathbf{x}, t) - \frac{1}{2}g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})]dt, \tag{3.3}$$

where $f(\mathbf{x}, t)$ and $g(t)$ are pre-defined schedule coefficients. This PF-ODE can be solved deterministically [61], mapping a noise sample to its corresponding data sample through the reverse process and the opposite through the forward process (inversion). This cycle-consistent conversion between image and latent representations is important in establishing dual diffusion implicit bridges.

Dual Diffusion Implicit Bridges. Dual Diffusion Implicit Bridges (DDIBs) [63] compose a diffusion inversion and generation process for solving image-to-image translation problems without requiring a paired image dataset. Instead, DDIBs use two diffusion models trained on different domains (or, analogously, one model with two different text conditions). DDIB inverts the source image into a noise latent via the forward PF-ODE and then decodes the latent in the target domain via the reverse PF-ODE. DDIBs can be interpreted as a concatenation of the Schrödinger Bridges from source-to-latent and latent-to-target, hence the dual bridges in its name. DDIBs enable solving transport between two distributions using a single pre-trained diffusion model. We build on this insight in an optimization context.

Optimization with Diffusion Model Approximates a Dual Schrödinger Bridge

Many generative vision tasks involve optimizing corrupted images to the image manifold. For example, in 3D generation, a 3D representation like NeRF is optimized to render natural images matching a prescribed text prompt. Methods like SDS enable this by using a pre-trained diffusion model as a prior. We propose formulating such optimization problems as solutions to an instantiation of a Schrödinger Bridges Problem (SBP). SBP finds cost-optimal paths between a source image distribution p_{src} and a target image distribution p_{tgt} [66, 14]. Optimizing a parametrized image toward the natural image distribution can be cast as finding the optimal paths between the current optimized image(s) and the natural image distribution. Instead of solving this problem directly, which would require training a generative model from scratch [39, 14, 10], we show that pre-trained diffusion models can be exploited as an optimizer that approximates the path. Further, the gradient computed by the existing score distillation methods can be viewed as the first-order approximation of this path. This formulation is illustrated in Figure 3.1

Let $\mathbf{x}_\theta \in \mathbb{R}^d$ represent a parametric image, *i.e.*, an image produced differentially by a model with parameter θ , such as a NeRF. To leverage the pretrained diffusion model, we add noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to obtain a latent at timestep t :

$$\mathbf{x}_{\theta,t} = \alpha_t \mathbf{x} + \sigma_t \epsilon$$

Suppose that $\psi_{t',\text{src}}$ and $\psi_{t',\text{tgt}}$ denote the paths obtained by solving the PF ODE as in Eq. 3.3 from t to 0, both starting from $\mathbf{x}_{\theta,t}$, such that $\psi_{0,\text{src}} \in p_{\text{src}}$, $\psi_{0,\text{tgt}} \in p_{\text{tgt}}$, $\psi_{t,\text{src}} = \psi_{t,\text{tgt}} = \mathbf{x}_{\theta,t}$. This forms a dual diffusion bridge [63] from $\psi_{0,\text{src}}$ to $\psi_{0,\text{tgt}}$. We approximate this path *per-iteration* using a pretrained diffusion model. We denote the displacement of this path as:

$$\epsilon_{\text{SBP}}^* = \psi_{0,\text{tgt}} - \psi_{0,\text{src}}. \quad (3.4)$$

Fully simulating this bridge involves solving two PF ODEs, which invokes dozens of neural function evaluations (NFEs) to estimate the gradient of each iteration. Instead, one can estimate each half of the bridge with a single-step prediction by computing two denoising directions $\epsilon_{\phi,\text{src}}$ and $\epsilon_{\phi,\text{tgt}}$. We thus obtain a first-order approximation of a dual diffusion bridge with the difference vector:

$$\epsilon_{\text{SBP}} = \epsilon_{\phi,\text{tgt}} - \epsilon_{\phi,\text{src}}, \quad (3.5)$$

which is subject to the following sources of errors.

1. **First-order approximation error.** Instead of performing full PF-ODE simulations, the single-step noising and prediction are less accurate and can induce errors. Recent work ISM [35] can be interpreted as reducing this error with a multi-step simulation to obtain $\mathbf{x}_{\theta,t}$.

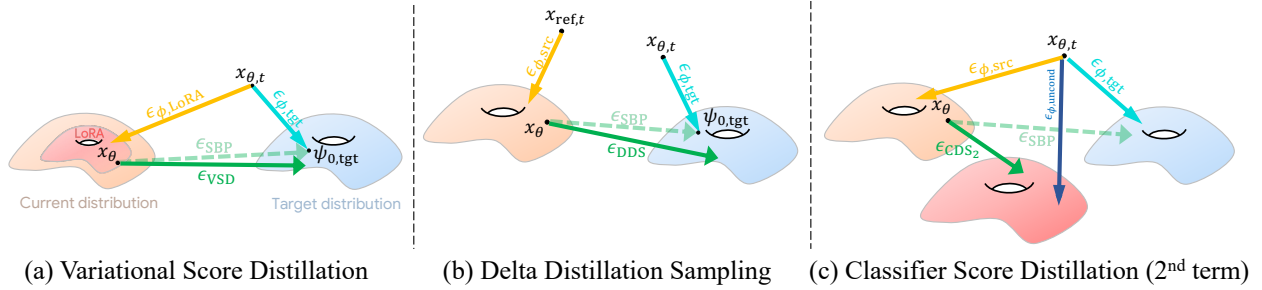


Figure 3.2: **Comparison of SDS variants under our analysis.** We illustrate the major gradient components of different SDS variants and provide a straightforward comparison with ϵ_{SBD} .

2. **Source distribution mismatch.** The dual diffusion bridge relies on $\epsilon_{\phi, src}$ accurately estimating the distribution of the current sample, \mathbf{x}_θ . A series of works can be viewed as improving this error [68, 28, 74] by computing more accurate $\epsilon_{\phi, src}$.

We show that $\epsilon_{\phi, tgt} - \epsilon_{\phi, src}$ is an effective gradient when both the source and target distribution are well expressed. Next, we discuss the popular score distillation methods under this analysis. We argue that their characteristic artifacts can largely be understood due to the errors above.

Analyzing Existing Score Distillation Methods

We analyze SDS and its variants through our framework by inspecting each component in the computed gradient. For notation, y_{tgt} is the text prompt representing the target distribution, and \emptyset denotes the unconditional prompt. For each method, we present its gradient update and discuss its implications.

Score Distillation Sampling [46]:

$$\epsilon_{SDS} = \epsilon_\phi(\mathbf{x}_{\theta,t}; \emptyset, t) + s \cdot (\epsilon_\phi(\mathbf{x}_{\theta,t}; y_{tgt}, t) - \epsilon_\phi(\mathbf{x}_{\theta,t}; \emptyset, t)) - \epsilon,$$

where s is the strength of classifier-free guidance. When s is small, the ϵ functions as an averaging term to regress the image to the mean. However, the SDS gradient has been shown to work best with extreme values of classifier-free guidance s like 100. We can rewrite the gradient to emphasize how the conditional-unconditional delta dominates at high CFG scales.

$$\epsilon_{SDS} = \underbrace{s \cdot (\epsilon_\phi(\mathbf{x}_{\theta,t}; y_{tgt}, t) - \epsilon_\phi(\mathbf{x}_{\theta,t}; \emptyset, t))}_{\text{Dominant when } s \gg 1} + \epsilon_\phi(\mathbf{x}_{\theta,t}; \emptyset, t) - \epsilon,$$

Experimentally, we produce very similar results at high CFG with or without the non-dominant terms. We argue that SDS should be interpreted through the dominant term,

which fits within our analysis. Under this interpretation, the unconditional distribution approximates the distribution of \mathbf{x}_θ poorly, instead representing images of any identity with low contrast and geometric artifacts. Figure 3.1(b) illustrates the effect of a poor approximation. The bridge from the unconditional to conditional distribution leads to the characteristic oversaturation and smoothing of SDS results.

Delta Distillation Sampling [20]:

$$\epsilon_{\text{DDS}} = \epsilon_\phi(\mathbf{x}_{\theta,t}; y_{\text{tgt}}, t) - \epsilon_\phi(\mathbf{x}_{\text{ref},t}; y_{\text{src}}, t),$$

where $\mathbf{x}_{\text{ref},t}$ is a noised version of a reference image in the image editing task. As shown in Figure 3.2 (b), this increases the *source distribution mismatch* since $\epsilon_{\phi, \text{src}}$ is not calculated based on the current optimized image $\mathbf{x}_{\theta,t}$.

Noise Free Score Distillation [28]:

$$\epsilon_{\text{NFSD}} = (\epsilon_\phi(\mathbf{x}_{\theta,t}; \emptyset, t) - (t < 0.2) \cdot \epsilon_\phi(\mathbf{x}_{\theta,t}; y_{\text{neg}}, t)) + s \cdot (\epsilon_\phi(\mathbf{x}_{\theta,t}; y_{\text{tgt}}, t) - \epsilon_\phi(\mathbf{x}_{\theta,t}; \emptyset, t)),$$

where the strength of classifier-free guidance s is set to 7.5 and y_{neg} = “unrealistic, blurry, low quality ...”. NFSD greatly reduces the guidance strength while it is observed to perform very similarly to SDS in practice. We can better explain this phenomenon since the prompt y_{neg} does not accurately describe the source distribution as it omits the image’s content. In addition, the second component with weight $s = 7.5$ still forms the major part of the gradient, which is the dominant term in SDS.

Classifier Score Distillation [74]:

$$\epsilon_{\text{CSD}} = w_1 \cdot (\epsilon_\phi(\mathbf{x}_{\theta,t}; y_{\text{tgt}}, t) - \epsilon_\phi(\mathbf{x}_{\theta,t}; \emptyset, t)) + w_2 \cdot (\epsilon_\phi(\mathbf{x}_{\theta,t}; \emptyset, t) - \epsilon_\phi(\mathbf{x}_{\theta,t}; y_{\text{src}}, t)),$$

where w_1 and w_2 are hyperparameters. As shown in Figure 3.2 (c), the second term approximates the bridge from the source distribution to the unconditional distribution, which is not ideal since it does not point to the target distribution. It explains the observation made by the authors [74] that this undermines the alignment with the text prompt. Therefore, the authors always anneal w_2 to 0 during the optimization. However, we show this often reintroduces the SDS artifacts in practice.

Variational Score Distillation [lee2024dreamflow, 68]:

$$\epsilon_{\text{VSD}} = \epsilon_\phi(\mathbf{x}_{\theta,t}; \emptyset, t) + s \cdot (\epsilon_\phi(\mathbf{x}_{\theta,t}; y_{\text{tgt}}, t) - \epsilon_\phi(\mathbf{x}_{\theta,t}; \emptyset, t)) - \epsilon_{\text{LoRA}}(\mathbf{x}_{\theta,t}; y_{\text{tgt}}, t).$$

Out of all the discussed methods, VSD attempts to minimize the *source distribution mismatch* error most directly by test-time finetuning a copy of the diffusion model with LoRA on the current set of \mathbf{x}_θ . Note that in the original paper, the use of LoRA was motivated based on a particle-based variational framework. Our analysis enables an alternative understanding of VSD. As shown in Figure 3.2 a), this approach is well-justified in our dual diffusion bridge framework. However, training a LoRA *every iteration* is computationally expensive, adds complexity, and introduces its own low-rank approximation errors. Given this insight, we propose a simple yet efficient approach to mitigating source distribution without LoRA.

Mitigating Source Distribution Mismatch with Textual Descriptions

Our analysis reveals that the LoRA model in VSD most closely approximates the distribution of the current optimized parametrized image, addressing the distribution mismatch error. Unfortunately, it incurs 200 – 300% runtime overhead on top of SDS, making it impractical, despite its significant performance gains. With this understanding, we propose a simple approach that better expresses the source distribution. Our insight is that pre-trained diffusion models have learned the distribution of natural and corrupted images through a combination of powerful text representation and enormous image-caption datasets. We find that by simply describing image corruptions with a text prompt, we can improve our estimate of the source distribution.

Specifically, we propose to use the gradient

$$\epsilon_{\text{ours}} = w \cdot (\epsilon_{\phi}(\mathbf{x}_{\theta,t}; y_{\text{tgt}}, t) - \epsilon_{\phi}(\mathbf{x}_{\theta,t}; y_{\text{src}}, t)),$$

where we get y_{src} by adding descriptions of the current image distribution to y_{tgt} (the base prompt). The remaining question is how to set this description. In generation tasks, we propose a simple two-stage solution.

1. We use ϵ_{SDS} to produce a generation with the method’s characteristic artifacts:
2. We switch to optimization with our gradient, ϵ_{ours} , to transport the image parameter toward the natural image distribution.

To describe the artifacts produced by SDS, we append the descriptors “, **oversaturated, smooth, pixelated, cartoon, foggy, hazy, blurry, bad structure, noisy, malformed**” and drop the descriptors of the high-quality generation. Note that in all of our generation experiments, the description of y_{src} is fixed as above. We explored searching for other prompts but did not find that variations in these descriptions made a big difference.

In editing tasks, we have an initialization that y_{src} describes accurately. In such cases, we omit the first SDS stage and only apply our gradient to optimization. We also append a “domain descriptor.” For instance, in painting-to-real, this is simply “, painting” to represent the initial distribution.

While the use of such negative prompting has been explored before, such as in NFSD, our analysis motivates a principled way to incorporate it into score distillation. We find that these simple modifications significantly narrow the quality gap between SDS and resource-intensive methods like VSD. We verify this finding experimentally with qualitative results and quantitative comparisons across applicable tasks.



Figure 3.3: **Text-to-image generation results with COCO Captions.** We compare different score distillation methods for generating images with COCO captions by optimizing a randomly initialized image. DDIM sampling indicates the lower bound that the diffusion model can achieve. VSD [68] and our method generate the least color artifacts while ours is more efficient than VSD.

3.1 Experiments

In this section, we test our proposed method on several generation problems where SDS is adopted. We compare against SDS and other task-specific baselines. Note that our goal is not to show another state-of-the-art text-to-3D generation method, but to verify our findings, where the proposed score distillation approach based on textual description efficiently improves the results by mitigating the source distribution mismatch error. We first perform a thorough experiment in a controlled setting on zero-shot text-to-image generation. Then, we compare it on text-guided NeRF optimization to SDS and VSD and evaluate the painting-to-real image translation task against image editing baselines. Please see more qualitative results, as well our method’s application to optical illusion generation and 3D-sketch-to-real task, in the appendix.

Zero-Shot Text-to-Image Generation with Score Distillation

To verify our analysis of existing SDS variants and the proposed method, we perform text-to-image generation by optimizing an image of size $64 \times 64 \times 4$ in the Stable Diffusion latent space [68, 28]. The benefit of choosing image generation as the evaluation task is that its generation quality has the least confounding variables among other tasks. (*e.g.*, in text-to-

Table 3.1: **Zero-shot FID comparison with different score distillation methods.** We report FID scores of text-to-image generation using 5K captions randomly sampled from the COCO dataset. The best score distillation result is indicated in **bold**, while the second best is underlined.

	DDIM (lower bound)	SDS [46]	NFSD [28]	CSD [74]	VSD [68]	Ours
Zero-Shot FID (\downarrow)	49.12	86.02	91.70	89.96	59.22	<u>67.89</u>
Zero-Shot CLIP FID (\downarrow)	16.56	28.39	29.25	27.07	18.86	<u>20.31</u>
Time per Sample (mins)	0.05	4.48	7.20	<u>6.21</u>	16.02	4.48

3D, many designs like regularizations [75], initialization [36], 3D representations [9, 65, 72, 64], and 2D prior models [55, 41, 40, 47, 76] could affect the final quality.)

We use the MS-COCO [37] dataset for the evaluation. Consistent with the prior study [3], we randomly sample 5K captions from the COCO validation set as conditions for generating images. For each caption, we optimize a randomly initialized the image with the score distillation gradients. We compare our method with several SDS variants including SDS [46], NFSD [28], CSD [74], and VSD [68]. For all the methods, we use the same learning rate of 0.01 and optimize for 2,500 steps where we generally observe convergence. We compute the zero-shot FID [21] and CLIP FID scores [31] between these generated images and the ground truth images. We also report results generated by DDIM with 20 steps as a lower bound for reference.

We report the FID scores and the time to optimize one image in Table 3.1. Among all the score distillation methods, VSD [68] achieves the lowest FID scores. However, it requires training a LoRA along the optimization process. Instead, ours achieves a comparable FID score with over $3\times$ faster speed. We visualize random examples generated by different score distillation methods in Figure 3.3. We notice that SDS and NSFSD suffer from the over-saturation and over-smoothness issues. CDS has slightly fewer color artifacts. VSD and ours generate the samples that most closely resemble the DDIM sampling.

Text-guided NeRF Optimization

We now evaluate the text-to-3D generation problem, where we intentionally aim to exclude variables that could affect the generation quality other than the score distillation methods. We use the ThreeStudio [19] repository to optimize a NeRF with settings tuned for Prolific-Dreamer stage 1 (NeRF optimization) [68]. Note that we do not perform stages 2 and 3, *i.e.* geometry fine-tuning and texture refinement. Specifically, we initialize the NeRF with the method proposed by Magic3D [36], use the regularization losses on the sparsity and opacity, and optimize for 25K steps. We adopt the native SDS and VSD guidance implementations for comparison.

We first show visual comparisons of different score distillation methods in Figure 3.4. We

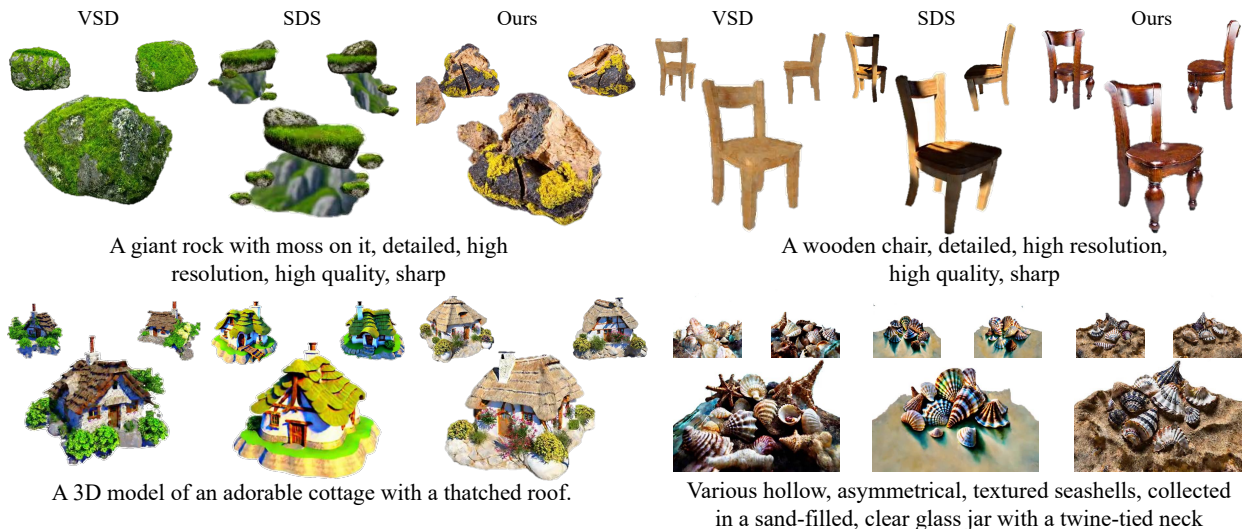


Figure 3.4: **Text-guided NeRF optimization with different score distillation methods.** We make a fair comparison of SDS and VSD for text-to-3D generation. For each generation, we show three uniformly sampled views. SDS results like the cottage and pepper mill still suffer from over-saturation problems, while ours and VSD can produce realistic details, color, and texture.

	ViT-L/14	ViT-B/16	ViT-B/32
SDS [46]	0.2811	0.3196	0.3139
VSD [68]	0.2837	0.3292	0.3166
Ours	0.2848	0.3282	0.3148

Table 3.2: **Quantitative comparisons of NeRF optimization.** We measure the average CLIP similarity of rendered views using SDS, VSD and our experimental method.

notice that SDS tends to generate fewer details, as shown by the rock and chair examples, and sometimes suffers from over-saturation issues, as in 2D, as demonstrated by the cottage and seashell examples. Instead, both VSD and ours can generate highly photo-realistic 3D objects, while ours does not require training a LoRA model and shares a similar computational cost as SDS.

We also perform a quantitative evaluation and user study on the NeRFs optimized based on 31 different text prompts. Note that this number is similar to the choice of existing works on the text-to-3D task [35, 32, 15]. However, different from these works that ignore the confounding 3D variables that contribute to the generation quality, we disentangle this by isolating the score distillation method as the only comparison variable. We follow these works to evaluate the generation quality with CLIP [48]. We report the CLIP similarity

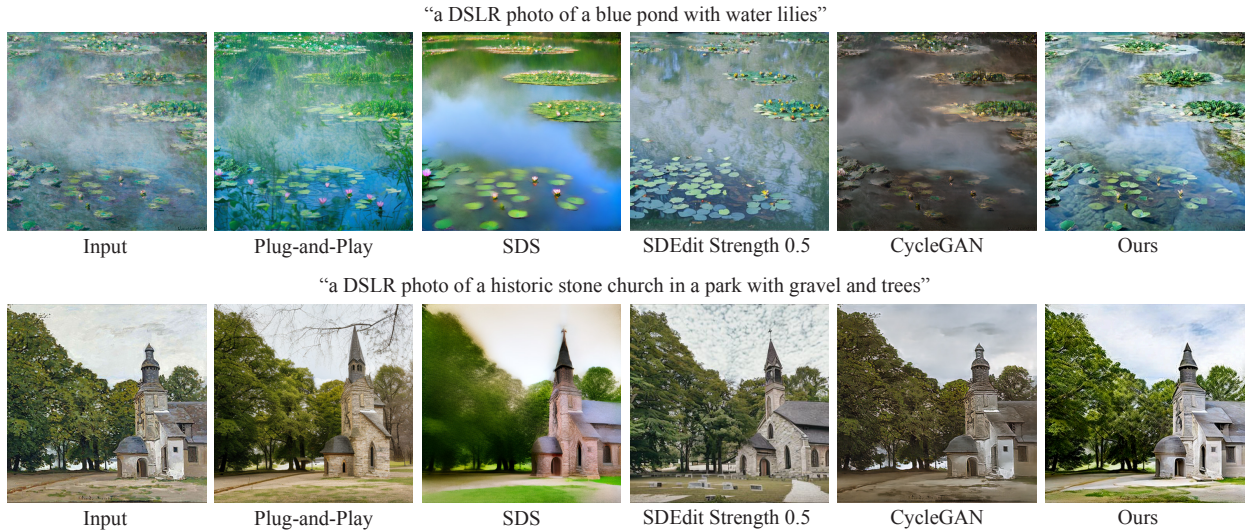


Figure 3.5: **Painting-to-Real comparison.** We compare our gradient in optimization to image restoration and image-conditional generation baselines. While SDEdit produces convincing textures, its difficult to find a strength value that balances structure and quality. Other baselines fail to reproduce natural image quality, while our method produces the best combination of quality and faithfulness.

in Table 3.2. Our method consistently outperforms SDS and achieves comparable results with VSD. In addition, in a user study consisting of 37 users, shown pairwise comparisons of rotating 3D renders (*i.e.*, comparisons of our result and a random choice of VSD or SDS, with the prompt: “For a text-to-3D system, given the prompt $[p]$, which result would you be happiest with?”), our results were chosen in 75.7% of all responses. We also show more results in the Appendix.

Painting-to-Real.

We examine our method’s ability to serve as a general-purpose realism prior. Paintings are “near-manifold” images, meaning they do not possess natural image statistics but live near the image distribution in image space. An effective image prior should guide a painting toward a nearby natural image through optimization.

We initialize a latent image by encoding scans of the artwork through Stable Diffusion’s encoder. We specify a prompt for each painting to condition the diffusion model and then apply the second optimization stage of our method (SDS stage omitted). We experimented with automatically generating prompts via pretrained vision language models but found the results inconsistent, so we leave this to future work. Since the large image datasets used to train diffusion models contain artwork, we append the domain descriptor “, painting” to y_{src} to optimize away from this distribution.

While SDS is proposed to leverage a pretrained text-to-image diffusion model as an image prior, its artifacts make it ineffective in practice. In comparison, our method realistically synthesizes details and relights the image naturally. We observe that SDS methods diverge more easily in 2D experiments than in 3D but that the issue can be mostly resolved with tuning. A future goal is to formulate a gradient that can be applied idempotently [56]. We compare with image reconstruction baselines in Figure 3.5 and provide a small gallery of painting-to-real results in Figure 3.6.

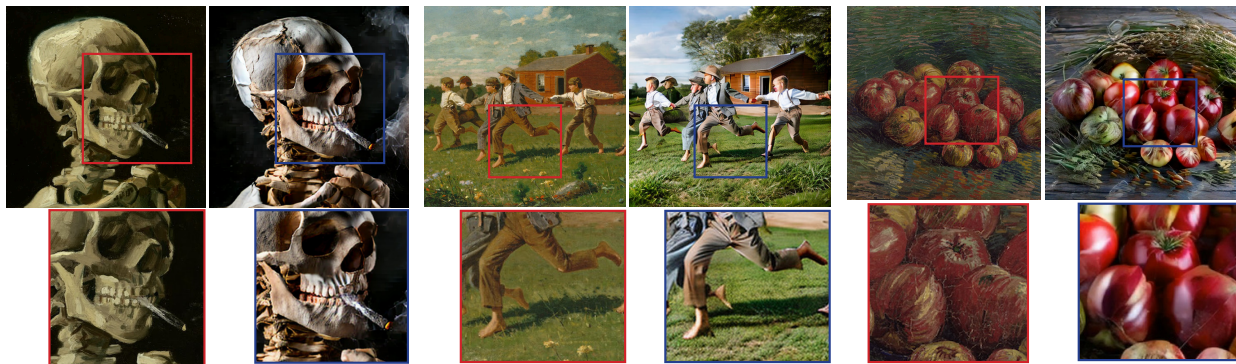


Figure 3.6: **Painting-to-Real results.** We show selected Painting-to-Real samples with diverse art styles and subjects. Initialization images are shown on the left, optimized images are shown on the right.

Chapter 4

Conclusion

4.1 Discussion

As we have shown that reducing the distribution mismatching error can significantly improve the generation quality of the score distillation optimization, it is natural to ask whether one can also reduce the first approximation error, induced by linear bridge estimation, to improve the results further. Several recent studies including SDI [42] and ISM [35] can be viewed as mitigating this error by replacing the single-step estimation with multi-step estimation. Instead of performing multiple PF-ODE steps, one can solve the entire PF-ODE path to recover the dual bridge and estimate the endpoint of the bridge $\psi_{0,\text{tgt}}$ that is coupled with $\psi_{0,\text{src}}$. In this way, we obtain the most accurate gradient direction with little approximation error $\epsilon_{\text{SBP}}^* = w \cdot (\psi_{0,\text{tgt}} - \psi_{0,\text{src}})$.

However, solving the inversion ODE is not trivial [27]. We noticed that the inversion can exaggerate the distribution mismatch error and cause the optimization to get stuck at a local optimal at the beginning of the optimization. Instead, the high variance of the single-step methods often shows more robustness to the input image. Therefore, we first perform the single-step score distillation optimization to obtain reasonable results before moving to solving the full bridge. We find that in text-to-2D, such a method can produce high-quality results closer to the DDIM sampling results, as demonstrated by a COCO-FID score of 55.65, which is better than VSD results. However, the same trend does not fully transfer to the text-to-3D experiments. We observe that it typically introduces additional artifacts and makes the optimization less stable. We leave the best way of leveraging this gradient as a future research exploration.

4.2 Closing Remarks

We present an analysis that formulates the use of a pre-trained diffusion model in an optimization framework as seeking an optimal transport between two distributions. Under this lens, we analyze SDS variants with a unified framework. We also develop a simple approach based on textual descriptions that work comparably well to the best-performing approach, VSD, without its significant computational burden. However, neither approach has yet to achieve the quality and diversity of images generated by the reverse process. We hope that our analysis enables the development of a more sophisticated solution that can one day achieve the same quality and diversity as the reverse process in an optimization framework. Combining our proposed method with multi-step approximations like ISM [35] or schedules like DreamFlow [32] could mitigate the first-order approximation error and further improve the efficiency, which is an interesting future research direction. With the rise of high-quality video diffusion models, we anticipate that the question of how to effectively use such models as a prior in various problems will become even more important.

Potential Social Impacts We analyze how to use a pre-trained image diffusion as a prior in an optimization setup, necessary for domains such as 3D. On the positive side, these models can empower individuals to make 3D content creation more accessibly without requiring specialized skills. Additionally, professional artists and designers could rapidly prototype and visualize their ideas, accelerating the creative process. On the negative side, the ease of generating visual content could facilitate the spread of misinformation, proliferate biases in the training set and enable the usage of generated content for malicious purposes. In addition, there are ethical concerns regarding the potential for job displacement in industries reliant on traditional art-making skills and the copyright issues appeared in the training dataset.

Bibliography

- [1] Thiemo Alldieck, Nikos Kolotouros, and Cristian Sminchisescu. *Score Distillation Sampling with Learned Manifold Corrective*. 2024. arXiv: 2401.05293 [cs.CV].
- [2] Sherwin Bahmani et al. “4D-fy: Text-to-4D Generation Using Hybrid Score Distillation Sampling”. In: *arXiv preprint arXiv:2311.17984* (2023).
- [3] Yogesh Balaji et al. “eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers”. In: *arXiv preprint arXiv:2211.01324* (2022).
- [4] Andreas Blattmann et al. “Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models”. In: *CVPR*. 2023.
- [5] Ryan Burgert et al. *Diffusion Illusions: Hiding Images in Plain Sight*. June 2023.
- [6] Tianshi Cao et al. “TexFusion: Synthesizing 3D Textures with Text-Guided Image Diffusion Models”. In: *ICCV*. 2023.
- [7] Dave Zhenyu Chen et al. “Scenetex: High-quality texture synthesis for indoor scenes via diffusion priors”. In: *arXiv preprint arXiv:2311.17261* (2023).
- [8] Dave Zhenyu Chen et al. “Text2tex: Text-driven texture synthesis via diffusion models”. In: *arXiv preprint arXiv:2303.11396* (2023).
- [9] Rui Chen et al. “Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023.
- [10] Tianrong Chen, Guan-Horng Liu, and Evangelos A Theodorou. “Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory”. In: 2022.
- [11] Yongxin Chen and Tryphon Georgiou. “Stochastic Bridges of Linear Systems”. In: *IEEE Transactions on Automatic Control* 61.2 (2016), pp. 526–531.
- [12] Yongxin Chen, Tryphon T. Georgiou, and Michele Pavon. “On the Relation Between Optimal Transport and Schrödinger Bridges: A Stochastic Control Viewpoint”. In: *Journal of Optimization Theory and Applications* 169 (2014), pp. 671–691. URL: <https://api.semanticscholar.org/CorpusID:8968928>.
- [13] Cheng Chi et al. “Diffusion Policy: Visuomotor Policy Learning via Action Diffusion”. In: *Proceedings of Robotics: Science and Systems (RSS)*. 2023.

- [14] Valentin De Bortoli et al. “Diffusion schrödinger bridge with applications to score-based generative modeling”. In: 2021.
- [15] Wenqi Dong et al. “Coin3D: Controllable and Interactive 3D Assets Generation with Proxy-Guided Conditioning”. In: *arXiv preprint arXiv:2405.08054* (2024).
- [16] Kevin Frans, Lisa Soros, and Olaf Witkowski. “Clipdraw: Exploring text-to-drawing synthesis through language-image encoders”. In: *Advances in Neural Information Processing Systems 35* (2022), pp. 5207–5218.
- [17] Daniel Geng, Inbum Park, and Andrew Owens. “Visual Anagrams: Generating Multi-View Optical Illusions with Diffusion Models”. In: *CVPR* (2024).
- [18] Shuyang Gu et al. “Vector quantized diffusion model for text-to-image synthesis”. In: *CVPR*. 2022, pp. 10696–10706.
- [19] Yuan-Chen Guo et al. *threestudio: A unified framework for 3D content generation*. <https://github.com/threestudio-project/threestudio>. 2023.
- [20] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. “Delta denoising score”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 2328–2337.
- [21] Martin Heusel et al. “Gans trained by a two time-scale update rule converge to a local nash equilibrium”. In: 2017.
- [22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *NeurIPS* 33 (2020), pp. 6840–6851.
- [23] Jonathan Ho et al. “Imagen video: High definition video generation with diffusion models”. In: *arXiv preprint arXiv:2210.02303* (2022).
- [24] Edward J Hu et al. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *ICLR*. 2022.
- [25] Shir Iluz et al. “Word-as-image for semantic typography”. In: *SIGGRAPH* (2023).
- [26] Ajay Jain et al. “Zero-shot text-guided object generation with dream fields”. In: *CVPR*. 2022, pp. 867–876.
- [27] Tero Karras et al. “Elucidating the Design Space of Diffusion-Based Generative Models”. In: *arXiv preprint arXiv:2206.00364* (2022).
- [28] Oren Katzir et al. “Noise-free score distillation”. In: *arXiv preprint arXiv:2310.17590* (2023).
- [29] Subin Kim et al. “Collaborative Score Distillation for Consistent Visual Editing”. In: 2023.
- [30] Juil Koo, Chanho Park, and Minhyuk Sung. “Posterior Distillation Sampling”. In: *arXiv preprint arXiv:2311.13831* (2023).
- [31] Tuomas Kynkäänniemi et al. “The Role of ImageNet Classes in Fréchet Inception Distance”. In: 2022.

- [32] Kyungmin Lee, Kihyuk Sohn, and Jinwoo Shin. *DreamFlow: High-Quality Text-to-3D Generation by Approximating Probability Flow*. 2024. arXiv: 2403.14966 [cs.CV]. URL: <https://arxiv.org/abs/2403.14966>.
- [33] Christian Léonard. *A survey of the Schrodinger problem and some of its connections with optimal transport*. 2013. arXiv: 1308.0215 [math.PR].
- [34] Tzu-Mao Li et al. “Differentiable vector graphics rasterization for editing and learning”. In: *ACM Transactions on Graphics (TOG)* 39.6 (2020), pp. 1–15.
- [35] Yixun Liang et al. “LucidDreamer: Towards High-Fidelity Text-to-3D Generation via Interval Score Matching”. In: *CVPR*. 2023.
- [36] Chen-Hsuan Lin et al. “Magic3D: High-Resolution Text-to-3D Content Creation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 300–309.
- [37] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: 2014.
- [38] Huan Ling et al. *Align Your Gaussians: Text-to-4D with Dynamic 3D Gaussians and Composed Diffusion Models*. 2024. arXiv: 2312.13763 [cs.CV].
- [39] Guan-Horng Liu et al. “I2SB: image-to-image Schrödinger bridge”. In: 2023.
- [40] Ruoshi Liu et al. “Zero-1-to-3: Zero-shot One Image to 3D Object”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 9298–9309.
- [41] Yuan Liu et al. “SyncDreamer: Generating Multiview-consistent Images from a Single-view Image”. In: *The Twelfth International Conference on Learning Representations*. 2024. URL: <https://openreview.net/forum?id=MN3yH2ovHb>.
- [42] Artem Lukoianov et al. “Score Distillation via Reparametrized DDIM”. In: *arXiv preprint arXiv:2405.15891* (2024).
- [43] Gal Metzger et al. “Latent-nerf for shape-guided generation of 3d shapes and textures”. In: *CVPR*. 2023, pp. 12663–12673.
- [44] Ben Mildenhall et al. “Nerf: Representing scenes as neural radiance fields for view synthesis”. In: *Communications of the ACM* 65.1 (2021), pp. 99–106.
- [45] Hyelin Nam et al. “Contrastive Denoising Score for Text-guided Latent Diffusion Image Editing”. In: *CVPR* (2022).
- [46] Ben Poole et al. “DreamFusion: Text-to-3D using 2D Diffusion”. In: 2023.
- [47] Guocheng Qian et al. “Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors”. In: *The Twelfth International Conference on Learning Representations*. 2024. URL: <https://openreview.net/forum?id=0jHkUDyE09>.
- [48] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *ICML*. 2021.

- [49] Aditya Ramesh et al. “Hierarchical text-conditional image generation with clip latents”. In: *arXiv preprint arXiv:2204.06125* (2022).
- [50] Aditya Ramesh et al. “Hierarchical text-conditional image generation with clip latents”. In: *arXiv preprint arXiv:2204.06125* (2022).
- [51] Aditya Ramesh et al. “Zero-Shot Text-to-Image Generation”. In: *ICML*. 2021.
- [52] Chitwan Saharia et al. “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding”. In: *arXiv preprint arXiv:2205.11487* (2022).
- [53] E. Schrodinger. “Sur la théorie relativiste de l’électron et l’interprétation de la mécanique quantique”. fre. In: *Annales de l’institut Henri Poincaré* 2.4 (1932), pp. 269–310. URL: <http://eudml.org/doc/78968>.
- [54] Christoph Schuhmann et al. “Laion-5b: An open large-scale dataset for training next generation image-text models”. In: *NeurIPS* (2022).
- [55] Yichun Shi et al. “MVDream: Multi-view Diffusion for 3D Generation”. In: *The Twelfth International Conference on Learning Representations*. 2024. URL: <https://openreview.net/forum?id=FUgrjq2pbB>.
- [56] Assaf Shocher et al. “Idempotent Generative Network”. In: 2024.
- [57] Uriel Singer et al. “Make-a-video: Text-to-video generation without text-video data”. In: *arXiv preprint arXiv:2209.14792* (2022).
- [58] Uriel Singer et al. “Text-to-4d dynamic scene generation”. In: *arXiv preprint arXiv:2301.11280* (2023).
- [59] Jascha Sohl-Dickstein et al. “Deep unsupervised learning using nonequilibrium thermodynamics”. In: 2015.
- [60] Jiaming Song, Chenlin Meng, and Stefano Ermon. “Denoising Diffusion Implicit Models”. In: *ICLR*. 2021.
- [61] Jiaming Song, Chenlin Meng, and Stefano Ermon. *Denoising Diffusion Implicit Models*. 2022. arXiv: 2010.02502 [cs.LG].
- [62] Yang Song et al. “Score-Based Generative Modeling through Stochastic Differential Equations”. In: *ICLR*. 2021.
- [63] Xuan Su et al. “Dual Diffusion Implicit Bridges for Image-to-Image Translation”. In: *ICLR*. 2022.
- [64] Jiaxiang Tang et al. *DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation*. 2023. arXiv: 2309.16653 [cs.CV].
- [65] Christina Tsalicoglou et al. “Textmesh: Generation of realistic 3d meshes from text prompts”. In: *arXiv preprint arXiv:2304.12439* (2023).
- [66] Gefei Wang et al. “Deep generative learning via schrödinger bridge”. In: 2021.

- [67] Haochen Wang et al. “Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation”. In: *CVPR*. June 2023, pp. 12619–12629.
- [68] Zhengyi Wang et al. “ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation”. In: *NeurIPS* (2023).
- [69] Tong Wu et al. “GPT-4V(ision) is a Human-Aligned Evaluator for Text-to-3D Generation”. In: 2024.
- [70] Ximing Xing et al. “DiffSketcher: Text Guided Vector Sketch Synthesis through Latent Diffusion Models”. In: 2023.
- [71] Ximing Xing et al. “SVGDreamer: Text Guided SVG Generation with Diffusion Model”. In: *arXiv preprint arXiv:2312.16476* (2023).
- [72] Taoran Yi et al. *GaussianDreamer: Fast Generation from Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models*. 2023. arXiv: 2310.08529 [cs.CV].
- [73] Kim Youwang, Tae-Hyun Oh, and Gerard Pons-Moll. “Paint-it: Text-to-Texture Synthesis via Deep Convolutional Texture Map Optimization and Physically-Based Rendering”. In: *CVPR*. 2024.
- [74] Xin Yu et al. *Text-to-3D with Classifier Score Distillation*. 2023. arXiv: 2310.19415 [cs.CV].
- [75] Junzhe Zhu and Peiye Zhuang. *HiFA: High-fidelity Text-to-3D Generation with Advanced Diffusion Guidance*. 2023. arXiv: 2305.18766 [cs.CV].
- [76] Zi-Xin Zou et al. *Sparse3D: Distilling Multiview-Consistent Diffusion for Object Reconstruction from Sparse Views*. 2023. arXiv: 2308.14078 [cs.CV].

Chapter 5

Appendix

5.1 Additional Experimental Setup

In this section, we describe our experimental setups in more detail.

Text-to-image generation with score distillation. For CSD, we follow the original paper [74] to use $w_1 = w_2 = 40$ at the initialization steps and anneal $w_2 = 0$ within the first 500 steps. We use $s = 100$ for SDS and $s = 7.5$ for NFSD and VSD, which are consistent with the best practice. We use $s = 40$ and $w = 25$ for our method. And we optimize with ϵ_{SDS} loss for 500 iterations and then switch to ϵ_{ours} for the rest of 2,000 iterations. For all the methods, we use a learning rate of 0.01, and we use a learning rate of $1e - 4$ to train the LoRA in VSD.

Text-guided NeRF optimization with score distillation. For our method, we optimize with ϵ_{SDS} loss for 20,000 iterations and then switch to ϵ_{ours} for the rest of 5,000 iterations. We use $s = 100$ and $w = 1$ for our method. We find that a high s is necessary to establish geometry in the first stage of the text-to-3D setting, but our method is not too sensitive to this hyperparameter in 2D. We use the rest of the learning rates and regularization strengths as the default settings.

5.2 More Visual Results

In this section, we provide extra visual results. Specifically, we show 3D sketch-to-real and optical illusion generation as additional applications of our method. We also report more comparisons and ablation studies of text-based NeRF optimization.

3D Sketch-to-Real Head-mounted displays with hand tracking are a natural platform for a sort of "3D sketching," where 3D primitives trail from your hand like ink from a pen. The resulting coarse mesh is structurally accurate but lacks geometric or texture detail. To



Figure 5.1: **3D sketch-to-real.** We introduce a conditional generation task in 3D where a coarse human-drawn mesh is optimized into a high-quality mesh. While SDS and our gradient both adhere to the prompt and shape conditions, our method produces higher fidelity colors and texture.

this end, we propose a new application that transfers these 3D sketches to more realistic versions. We extend our text-to-3D solution to generate these details.

We first fit an implicit SDF volume to multi-view renders of the mesh, then apply our gradient with the same schedule as in text-based NeRF optimization. We lower the learning rate for geometry parameters to prevent divergence from the guiding sketch. Holding other hyperparameters equal, we compare our gradient and the SDS gradient in Figure 5.1.

Illusion Generation. Prior works have shown that diffusion models can be leveraged to generate optical illusions [17, 5]. In these settings, the same image looks semantically different when transformed. To use the diffusion model sampling process, a previous study shows that the transformation has to be orthogonal [17]. However, there remain interesting illusions that are not formed by orthogonal transformation. One such is the rotation overlays. Given a base and a rotator image, by composing the base image with the rotator image at different angles, rotation overlays use two images to display four images. As such composition is not defined by an orthogonal matrix, the existing method [5] employs SDS to optimize the base and rotator images. Such a method suffers from the over-saturation problem, as shown in Figure 5.2. We show that our method can generate such optical illusions with better visual quality.

Additional text-guided NeRF optimization results. For text-guided NeRF optimization comparison against baselines, we follow show more results in Fig. 5.4. We test on the prompts used in the original paper [68] and additional prompts [69] that we find to be challenging. We notice that SDS often suffers from over-saturation problems. Our method does not require training a LoRA while it can still improve SDS by getting rid of the color artifacts and generating more details.

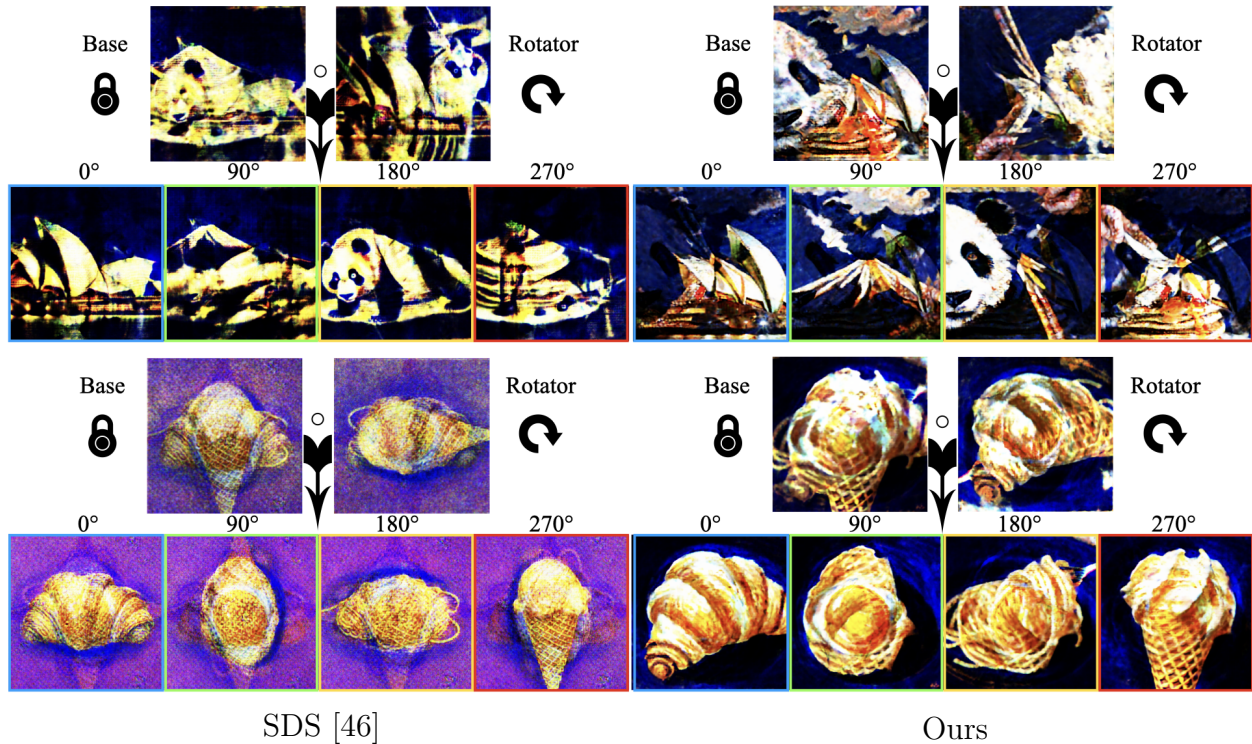


Figure 5.2: **Diffusion illusions.** We generate overlaid optic illusions with SDS and our method. While SDS suffers from color artifacts, our methods produce more details and proper color.

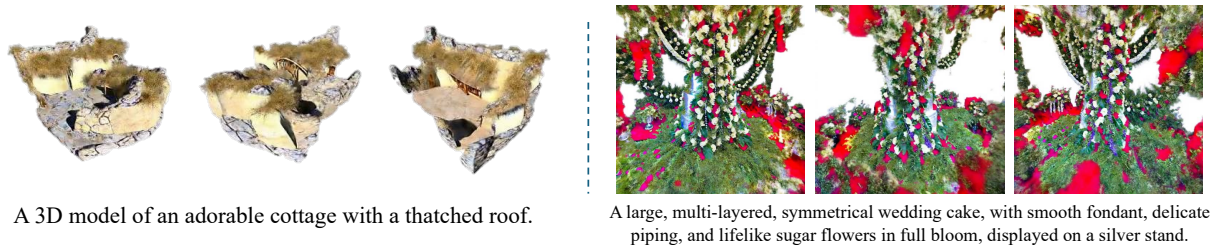


Figure 5.3: **Ablation study of our method without stage 1.** We show directly optimizing with y_{src} from the start could undermine the quality of the geometry and produce unnecessary content.

Ablation study of stage 2. Instead of switching to stage 2 during the optimization process, we ablate with starting without any SDS optimization from the beginning. That is, we always use the y_{src} with the descriptors “, oversaturated, smooth, pixelated, cartoon, foggy, hazy, blurry, bad structure, noisy, malformed”. As shown in Figure 5.3,

this makes it hard to generate the proper geometry even though the local texture looks reasonable and is inclined to produce excessive details that are not described by the texts. We suspect that this is because using y_{src} increases the mismatching error at the beginning of the optimization process when the initialization does not resemble the target prompt at all.



Figure 5.4: Additional comparison of text-guided NeRF optimization. We show more examples to compare with different distillation methods, SDS and VSD.