Birth and Death of a Rose

Chen Geng Yunzhi Zhang Shangzhe Wu Jiajun Wu
Stanford University

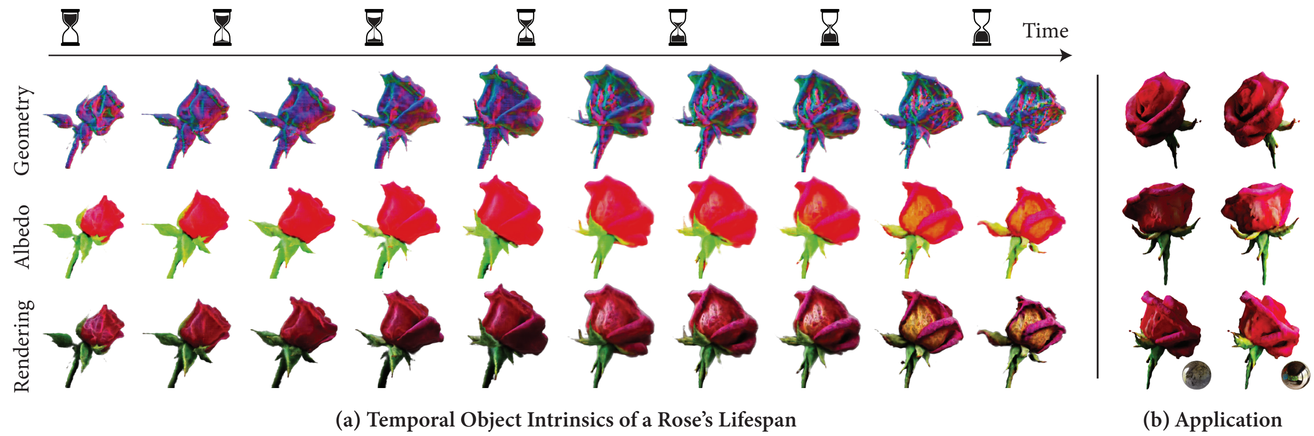

Figure 1. **The lifespan of a rose's *object intrinsics* generated with our method.** We present a pipeline to generate temporally evolving sequences of 3D objects' geometry and material properties, including albedo, roughness, and metallic parameters (only albedo is shown here), as shown in (a), by distilling from 2D foundation models without any 3D data. As shown in (b), the generated assets can be rendered from any viewpoints and lighting conditions (environment map shown in the last row). See the supplemental website for animations.

Abstract

We study the problem of generating temporal object intrinsics—temporally evolving sequences of object geometry, reflectance, and texture, such as a blooming rose—from pre-trained 2D foundation models. Unlike conventional 3D modeling and animation techniques that require extensive manual effort and expertise, we introduce a method that generates such assets with signals distilled from pre-trained 2D diffusion models. To ensure the temporal consistency of object intrinsics, we propose Neural Templates for temporal-state-guided distillation, derived automatically from image features from self-supervised learning. Our method can generate high-quality temporal object intrinsics for several natural phenomena and enable the sampling and controllable rendering of these dynamic objects from any viewpoint, under any environmental lighting conditions, at any time of their lifespan. Project website: `https://chen-geng.com/rose4d`.

> Those granted the gift of seeing more deeply can see beyond form, and concentrate on the wondrous aspect hiding behind every form, which is called life.
>
> — *Hilma af Klint*

1. Introduction

As illustrated in Figure 1, starting from a bud, a rose gradually expands its petals and blossoms into full glory, only to wither and conclude its lifespan. This inevitable and unidirectional evolution—a journey shared by all living creatures on Earth—relentlessly and chronologically transforms their *object intrinsics*: geometry, reflectance, and texture. Such progression of their object intrinsics characterizes our visual conception of the aging of an object, which we collectively refer to as its *temporal object intrinsics*.

Traditionally, creating graphics assets with realistic chronologically evolving object intrinsics requires extensive, object-specific manual effort and expertise [9, 18, 42]. Instead, we pursue a learning-based approach to generate such physically grounded graphics assets with no intervention. The generated temporal object intrinsics can be seen as a "3D time-lapse volumetric video", from which we can sample instances and render them from any viewpoint, under any lighting condition, and at any time of their lifespan.

It is challenging to generate temporal object intrinsics in a supervised manner due to the lack of annotated data. Therefore, we explore the potential of generative pipelines distilled from 2D diffusion models. Existing techniques like Score Distillation Sampling (SDS) [41] have shown promising results in 3D generation, but they are not directly transferable to our case of distilling temporal object intrinsics with significant changes in their geometry and texture. In 3D distillation, it is already well known that SDS-like methods struggle to maintain 3D consistency due to the lack of 3D information in the 2D diffusion model, often referred to as the Janus problem [1]. Unfortunately, the situation worsens in our optimization of 4D representations due to global inconsistency across both space and time: not only may a signature view appear from multiple camera viewpoints, but a common temporal state may also appear repeatedly throughout the entire duration.

To mitigate such 4D inconsistencies, we propose *Neural Templates* for temporal-state-conditioned distillation. Neural Template, a mapping that takes in viewpoint and time and outputs the "temporal state" information of modeled natural process, captures the lifespan of dynamic object's intrinsics; it can be automatically constructed by forging self-supervised image features [8, 39] obtained from a rough initial 4D reconstruction depicting the dynamic process. It allows us to anchor the distillation gradients to a particular viewpoint and timestamp by conditioning the diffusion model on a 2D Neural State Map representing the temporal state. This significantly improves the distillation efficiency and 4D consistency, as each view receives distillation signals tailored to that particular temporal state.

To model the photorealistic texture of an object, we further decompose its appearance into physically-based surface materials components and recover these representations during the distillation process using a differentiable PBR renderer. We also propose a hybrid 4D representation for consistent yet high-fidelity generations.

To summarize, we propose a new task of generating *temporal object intrinsics*, in the form of temporally-evolving sequences of 3D shape, reflectance, and texture. We introduce a framework to distill 4D-consistent temporal object intrinsics from pre-trained 2D diffusion models. Central to this framework is a canonical Neural Template that anchors the distillation signals to specific temporal states. We test this framework on several different object categories. We quantitatively compare its performance with prior state of the art on automatic 4D generation, suggesting advantages of the proposed methods across different examples. A further ablation study shows that the core modules and techniques proposed in this framework are crucial for performance.

2. Related Work

Dynamic Modeling and Generation. Understanding and further synthesizing the dynamic motion of objects have been a challenging and long-standing task in Computer Graphics. A prominent challenge is the limited amount of data due to the high cost of motion data collection. Prior works address this challenge via simulation-based techniques [9, 18, 42] or human-designed templates [32, 40], but the assumption of access to the simulator or template limits applying such methods to generic object categories. More recently, several methods have weakened the template assumption and required only the skeleton structure by introducing skinning-based representation, learning from pure image [28, 64, 70, 71] or video data [67] using pixel reconstruction as objectives, and are extended to the task of 4D generation [54]. These methods have been mostly applied to articulated objects or animals, while in comparison, our method can be applied to generic objects such as flowers, which require a highly expressive representation such as the proposed Neural Template.

Controlling and Distilling from Diffusion Models. One of the applications of modeling 4D intrinsics is to synthesize 4D videos of an object similar to the input, which can be cast into a controllable generation problem. Recent methods leverage pre-trained image [47] or video diffusion models [5] to support conditioning signals such as camera viewpoints [31], lighting conditions [11, 21, 25, 74], type of motions [81], and identity personalization [14, 48]. These methods typically train an additional lightweight network [77], low-rank layers [16], or explicitly distill the pretrained model into a new representation [12, 19, 35, 41]. Ours falls into the last category; compared with previous work that uses score distillation to generate 4D contents [2, 15, 19, 27, 29, 30, 44, 45, 55, 56, 62, 65, 68, 72, 73], our method supports the modeling of lighting, viewpoint, and timestep all at once and tackles the full task of 4D intrinsic modeling, while prior works tackle parts of the task.

Intrinsic Decomposition. Recovering object intrinsics, namely its geometry, texture, and material, is a long-standing task [4]. This information is not directly available in image or video observations, and previous methods tackle this task by optimizing for pixel reconstruction [3, 6, 20, 76, 78], learning with supervised learning targets using ground truth material from synthetic data [26], or generative modeling [10, 11, 21, 46, 57, 66, 74, 79, 80]. However, these methods exclusively focus on static intrinsics, while our method, in addition, accounts for the temporal evolutionary nature of these properties.

3. Overview

3.1. Task

Our goal is to generate *temporal object intrinsics*, a chronologically ordered sequence of the *object intrinsics* in a natural process, such as "*a rose blooming*", where "object intrinsics" refers to the intrinsic properties of an object—in our case, its geometry and material.

The natural processes considered in this paper are tem-

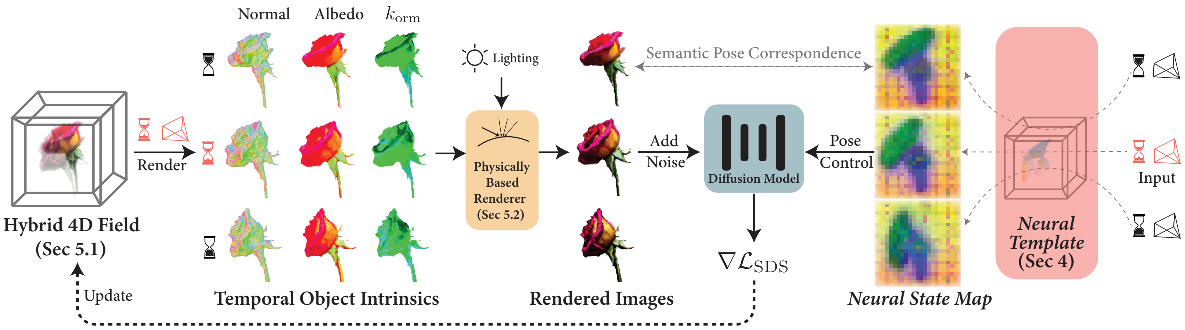

Figure 2. **Overview of the process to generate temporal object intrinsics.** For a modeled natural process, we first estimate a task-specific *Neural Template* to get a 4D-consistent representation of its temporal state changes (Sec. 4). Then we iteratively optimize a hybrid 4D intrinsics field (Sec. 5.1) to generate temporal object intrinsics. In each iteration of optimization, we sample a camera pose and timestamp to render this optimizable 4D field into intrinsic maps. These maps are subsequently rendered into RGB images using a physically-based renderer (Sec. 5.2). Concurrently, the same camera pose and timestamp are used to query the Neural Template for the corresponding Neural State Map. This map conditions a fine-tuned 2D diffusion model, which provides guidance signals to update the 4D representation. See Sec. 4 and Figure 4 for details on Neural Template.

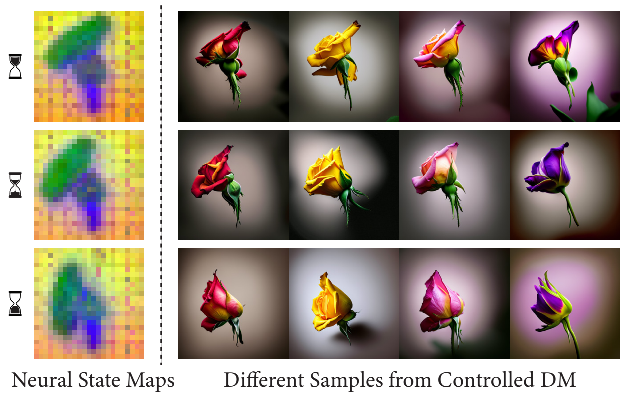

Figure 3. **A motivating example: *Neural State Map* as a conditional control signal for diffusion models.** Compressed DINOv2 features, or "Neural State Map", represent both the temporal state and spatial viewpoint information throughout a temporal natural process, and inject effective 4D control into 2D diffusion models.

porally evolving transformations imposed on the object by the external environment. Formally, these processes are characterized by several key properties: they span a long-range time frame, from hours to days; they are inevitable, driven by external factors; they involve significant changes in the object's geometry and/or material throughout its lifespan; and they occur in chronological sequence with a unidirectional order of temporal state changes.

3.2. Method Motivation

We propose to distill 4D temporal object intrinsics from 2D foundation models. Recent works have utilized techniques such as SDS [41, 58] to generate 4D content [2, 44]. However, these methods practically do not perform well on this task. On the one hand, they model appearance with view-dependent radiance, which may not be fully physically accurate and does not support relighting. More critically, as shown in our experiments (Figure 6), those methods typically generate minor and unrealistic motions, insufficient for representing the significant temporal state changes observed in objects undergoing drastic transformations over their lifespan.

For prior works, the primary challenge arises from the discrepancy between the spatial capacity of 2D diffusion models and the 4D information needed in the temporal object intrinsics. It is well-known that common score-distillation-based 3D generation methods [41, 58] often result in "Janus problem", where a signature view repeats itself in various sides of the generated asset. This is mainly because the 2D diffusion model has limited knowledge of the 3D viewpoint control. In the case of 4D generation, apart from repetitive views, the generated instances may exhibit repetitive *temporal states* across different timestamps, due to the lack of sufficient temporal anchoring.

Therefore, distilling from 2D diffusion models necessitates a clear 3D and temporal control signal [77] to enable high-fidelity 4D object intrinsics generation. What would be a good control signal? In traditional Computer Graphics literature [23, 24, 32], *skeletons* are commonly used to represent the motion state. As we are interested in generic objects beyond articulated ones, we need a more generic representation that shares a similar role as skeletons to encode semantic affinity across object parts.

In fact, after spatial downsampling and Principal Component Analysis (PCA), 2D image feature maps extracted using DINOv2 [39] can effectively provide such semantic affinity information, which is similar to the projection of skeleton in traditional animation. As shown in Figure 3, these feature maps effectively encode the temporal state as well as the viewpoint information for a natural process. We therefore call those maps "Neural State Maps". We empirically found that these maps can serve as conditional signals

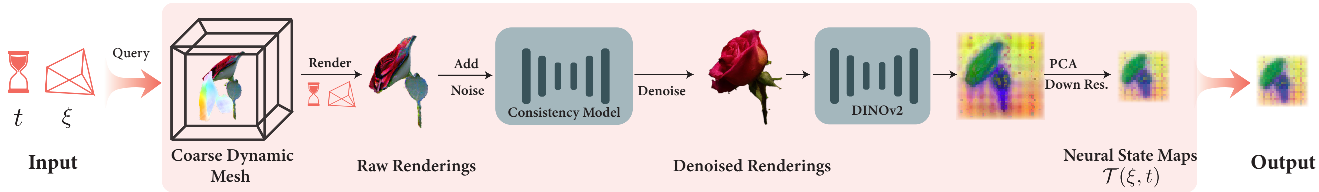

Figure 4. **Internal structure of Neural Template.** Taking camera pose ξ and timestamp t as input, the neural template first render a task-specific coarse dynamic mesh under the provided condition. It is further denoised with a consistency model and encoded to a feature map, which will be further used to control the diffusion models.

for ControlNet [77] to control the temporal state and viewpoint of generated images.

This finding bridges the 2D diffusion model and 4D information naturally. For a natural process, we can construct a canonical representation that tells us the temporal state information given a query camera viewpoint ξ and a timestamp t. We call this a "Neural Template", which defines a mapping $\mathcal{T}(\xi, t)$ from the camera pose and timestamp to the Neural State Map. We will further describe how we construct such mappings in Sec. 4.

3.3. System Overview

We build a system that takes in the prompt of a natural process, such as "*a rose blooming*", and generate temporal object intrinsics, as shown in Figure 2. The generation process is split into two stages. In the first stage, we construct a coarse dynamic mesh to represent the 4D temporal stages in the natural process. The constructed coarse dynamic mesh can be used to obtain the Neural Template as in Figure 4, which will be used in the second stage. We discuss the details of this process in Sec. 4.

In the second stage, we generate temporal object intrinsics by iteratively optimizing a hybrid 4D object intrinsics representation with gradients from 2D diffusion models. At each iteration, we randomly sample a camera viewpoint and a timestamp, and render an image with a physically-based renderer. The rendered image is then perturbed with noise and fed into a diffusion model conditioned on the Neural State Map, to obtain a generative score for updating the hybrid 4D field. This process is further discussed in Sec. 5.

4. Neural Template

As introduced in Sec. 3, the Neural Template aims to capture the canonical temporal state changes during the modeled process. Formally speaking, it is a mapping that takes a query viewpoint $\xi \in \mathbb{SE}(3)$ and a timestamp $t \in \{1, \cdots, T\}$, where T is the total number of timestamps, and outputs a corresponding "Neural State Map" $\mathcal{T}(\xi, t) \in \mathbb{R}^{H_F \times W_F \times d_F}$ (recall that it is a 2D feature map from DINOv2 after PCA and down-resolution).

The internal structure of Neural Template can be found in Figure 4. It consists of a task-specific coarse dynamic mesh, which stores coarse canonical pose information of the progressive state changes throughout the natural process, and several pre-trained frozen models that convert the renderings from the canonical RGB field to the final Neural State Map. We first discuss how we obtain the task-specific coarse dynamic mesh (Sec. 4.1), and then illustrate the algorithm to obtain the output Neural State Map (Sec. 4.2). More details of this section can be found at the Sec. A of the supplementary material.

4.1. Estimating Coarse 4D Geometry

For each natural process considered, we construct a coarse dynamic mesh to represent the semantic structure throughout the process. Such coarse dynamic mesh does not need to produce high rendering quality. Rather, it only needs to represent the coarse shape and appearance changes throughout the process.

We present one way to obtain such a dynamic mesh. Assuming that we have a 2D video diffusion model capable of modeling the motion of the natural process[1], we can sample a video V_{ref} from this model. We perform a physically-grounded 4D reconstruction of the video V_{ref} to obtain the dynamic mesh, using the following procedure.

We first choose one frame in the video as the canonical frame, and use off-the-shelf 3D reconstruction models [31, 59] to obtain a static 3D mesh as the representation of the static canonical frame. We then optimize a deformation field $D(\xi, t)$ that explicitly deform the 3D canonical mesh to match the geometry and appearance of other frames in the video. During each iteration, we rasterize the 3D scene flow into 2D flow maps and supervise them with optical flow estimated from off-the-shelf models [51]. We use additional regularization terms including As-Rigid-As-Possible (ARAP) [17, 53, 63] to ensure the physical plausibility. More details can be found in the Sec. A of the supplementary material.

4.2. Extracting Neural State Map

Given a camera pose ξ and a timestamp t, we render the aforementioned dynamic mesh into an RGB image. In order

[1] Practically, we train a LoRA [16] on a set of manually-collected time-lapse videos on pre-trained video diffusion models [69]. See supplementary material for details.

to use this image to control the diffusion model, we map it to a Neural State Map defined in Sec. 3.2, as described next.

Denoising raw renderings with consistency model. Existing 3D/4D reconstruction models typically have difficulties in accurately reconstructing rarely-seen natural objects, such as *roses*. Therefore, the coarse dynamic mesh obtained above might not align with the natural image distribution, on which the foundation model DINOv2 is trained.

To address this issue, we propose to use consistency distillation to aggregate priors from 2D generative models while adhering to the Neural State Map features. For any given pre-trained diffusion model, we use a consistency model [33, 52] distilled from it, denoted as $\hat{\epsilon}_\gamma$, to obtain the denoised samples of the noisy renderings. Let

$$\mathbf{f}_\gamma(\mathbf{z}, \mathbf{c}, \tau) = c_{\text{skip}}(\tau) \cdot \mathbf{z} + c_{\text{out}}(\tau) \left(\frac{\mathbf{z} - \sigma_\tau \hat{\epsilon}_\gamma(\mathbf{z}, \mathbf{c}, \tau)}{\alpha_\tau} \right), \tag{1}$$

where τ is the diffusion step, c_{out} and c_{skip} are functions satisfying certain boundary conditions [52], σ_τ, α_τ follow a diffusion noise schedule, $\mathbf{z}$ is the latent code with noise added at noise level τ, and $\mathbf{c}$ is the text prompt corresponding to the input video, *e.g.*, 'rose'.

Encoding RGB map into Neural State Map. After obtaining denoised image with this process, we use a pre-trained 2D self-supervised image encoder F [39] to obtain a feature map, which is further downsampled after a PCA step to encourage the Neural State Map to capture only low-frequency information.

5. Guided 4D Object Intrinsics Synthesis

The estimated Neural Template $\mathcal{T}$ from Sec. 4 can be used as guidance for generating a 4D instance $\mathcal{G}$. We first describe our 4D representation for $\mathcal{G}$ (Sec. 5.1) and its rendering process (Sec. 5.2), and then explain how to distill $\mathcal{G}$ via 2D diffusion models controlled with $\mathcal{T}$ (Sec. 5.3).

5.1. Hybrid 4D Representation

We now introduce an implicit neural 4D representation for $\mathcal{G}$ that is both expressive and temporally consistent. The geometry of $\mathcal{G}$ is modeled as a time-evolving implicit level-set [34, 38, 61], represented by an implicit neural SDF. Given a spatial coordinate $\mathbf{x}$ and timestamp t, we compute a feature vector $f_{\text{4D}}(\mathbf{x}, t)$, which is subsequently passed through two MLPs. These MLPs output SDF values and material parameters essential for neural rendering [36, 60] as detailed in Sec. 5.2.

The neural 4D feature vector f_{4D} is computed via a hybrid 4D representation with K-Planes [13] representing low-frequency information and Neural Graphical Primitives (NGP) [37] representing high-frequency details.

For K-Planes, we maintain 6 planes denoted with $\mathbf{P}_c$ to capture temporal information with tensorial decomposition. The low-frequency feature of K-Planes f_{low} is calculated with

$$f_{\text{low}}(\mathbf{x}, t) = \prod_{c \in C} \psi(\mathbf{P}_c, \pi_c(\mathbf{x}, t)), \tag{2}$$

where π_c projects $(\mathbf{x}, t)$ onto the c'th plane, ψ denotes interpolation, and Π denotes the Hadamard product.

We find that f_{low} represents temporal consistency information but struggles to represent high-frequency details. To address this, we represent the high-frequency details using several NGPs for K keyframes. For each of the keyframes, a multi-resolution hash grid is constructed. The feature for a rendered timestamp t is further obtained with

$$f_{\text{high}}(\mathbf{x}, t) = \textbf{lerp}\left(\frac{t - t_i}{t_{i+1} - t_i}, \psi_{t_i}(\mathbf{x}), \psi_{t_{i+1}}(\mathbf{x}) \right), t_i \le t \le t_{i+1}, \tag{3}$$

where ψ_t is the NGP encoding for timestamp t, and t_i and t_{i+1} are two keyframes around the sampled timestamp. The full 4D feature f_{4D} is computed by concatenating f_{low} and f_{high}.

5.2. Physically-Based Rendering

We now describe the rendering procedure for the hybrid representation of $\mathcal{G}$ from Sec. 5.1.

Shading. Following common conventions, we use the Disney Principled PBR material model [7] to perform the shading process. Specifically, the material parameters consist of diffuse color k_d, the visibility map $V(\mathbf{x})$, and $k_{\text{orm}} = (o, r, m)$ with roughness r, metallic parameter m, and o left unused. The specular color k_s is computed with $k_s = (1 - m) \cdot 0.04 + m \cdot k_d$. The outgoing radiance $L(\omega_o)$ in direction ω_o can be represented with the rendering equation [22]:

$$L(\mathbf{x}, \omega_o) = \int_\omega L_i(\mathbf{x}, \omega_i) f(\omega_i, \omega_o) V(\mathbf{x})(\omega_i \cdot \mathbf{n}) \mathrm{d}\omega_i, \tag{4}$$

where $\mathbf{n}$ is the surface normal, $L_i(\mathbf{x}, \omega_i)$ is the incoming radiance from direction ω_i, $f(\omega_i, \omega_o)$ is the BSDF computed from the material parameters.

Volume rendering. After computing the shading value at each spatial coordinate $\mathbf{x}$ and timestamp t, we follow a standard differentiable rendering process [36, 60] to produce the rendered frame at time t. These frames are then concatenated to form a complete video V_G. More details on rendering can be found in the supplementary material.

5.3. 4D Distillation with Neural State Map Controls

To obtain parameters θ of 4D content $\mathcal{G}$ using the above representation, we propose an optimization framework that distills from pre-trained 2D diffusion models [47], by adapting

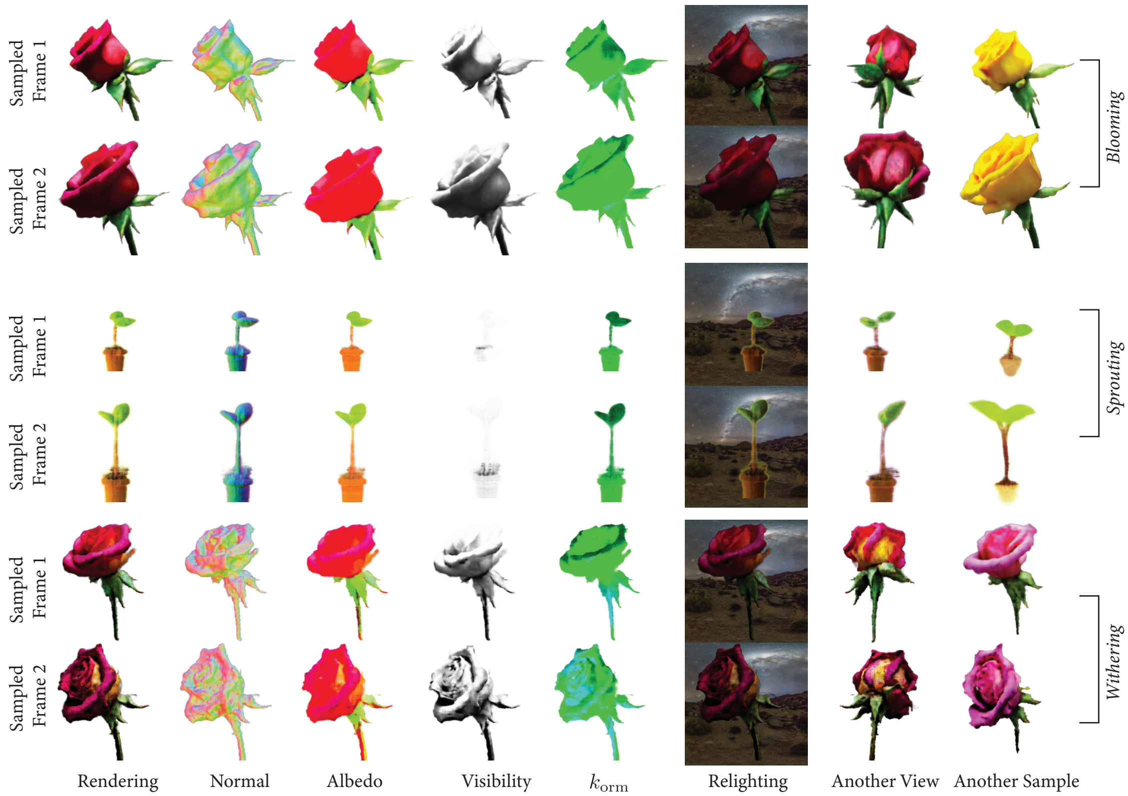

Figure 5. **Results on temporal object intrinsics generation.** We show the generated 4D sequence with our method on part of the evaluated phenomena. Our method synthesizes the geometry and albedo, as well as material properties of roughness and metallic, with high fidelity. The 4D sequence can be used for relighting and novel view synthesis. We also show another sample in the generation. Please check the supplemental material for more results.

these models to condition on Neural State Map computed from $\mathcal{T}$ as discussed in Sec. 3.

Specifically, to inject Neural State Map conditions, we follow ControlNet [77] and train backbone models on synthetic data. The training dataset consists of data pairs $(I, F(I))$, where I is an image generated from open-accessible APIs [49] and $F(I)$ is the control signal computed with neural pose encoder F from Sec. 3. The trained models are used to guide optimization of $\mathcal{G}$.

We optimize the representation in Sec. 5.1 with score distillation from Neural State Map-controlled diffusion models (Sec. 5.3). As each optimization iteration, we sample a camera viewpoint ξ and a timestamp t, and compute the following SDS gradient:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\theta; \xi, t, \mathbf{y}) = \mathbb{E}_{\tau,\epsilon,\xi,t}\left[w(\tau)\left(\hat{\epsilon}\left(\mathbf{z}_\tau; \tau, \mathcal{T}(\xi, t), \mathbf{y}\right) - \epsilon\right)\frac{\partial \mathbf{z}_\tau}{\partial \theta}\right], \quad (5)$$

where $w(\tau)$ is a weighting function, $\mathbf{y}$ is the text prompt, $\mathbf{z}_\tau$ is the latent code of image rendered with the physically-based renderer, with a noise level τ.

We also include a temporal regularization term to ensure the temporal consistency of the generated $\mathcal{G}$. We render a video $V_{\mathcal{G}}(\xi)$ with viewpoint ξ. It is then fed into a video diffusion model [69] with added noise and is denoised into a refined video $\hat{V}_{\mathcal{G}}(\xi)$ with a better temporal consistency. The refined version is used to supervise the representation with $\mathcal{L}_{\text{Vid}}(\theta; \xi) = ||V_{\mathcal{G}}(\xi) - \hat{V}_{\mathcal{G}}(\xi)||_2^2$. More details on the optimization can be found in the Sec. B.

	User Study			CLIP ↑
	Motion Alignment ↑	Motion Realism ↑	Visual Quality ↑	
DG4D [44]	2.26%	5.26%	11.28%	29.65
4D-Fy [2]	4.21%	11.13%	21.95%	30.11
STAG4D [75]	16.69%	13.83%	5.71%	28.01
Ours	**76.84%**	**69.77%**	**61.05%**	**30.83**

Table 1. **Comparison with 4D generation baselines.** We perform a user study on 7 different natural processes to assess the performance of our method and previous 4D generation methods in three aspects. We report the percentage of cases in which a method ranks first, which indicates a significant advantage of our method. We also report a CLIP score.

6. Experiments

We test the proposed method on different examples of dynamic natural phenomena, all with a temporal evolving nature of object intrinsics. We also compare our model with

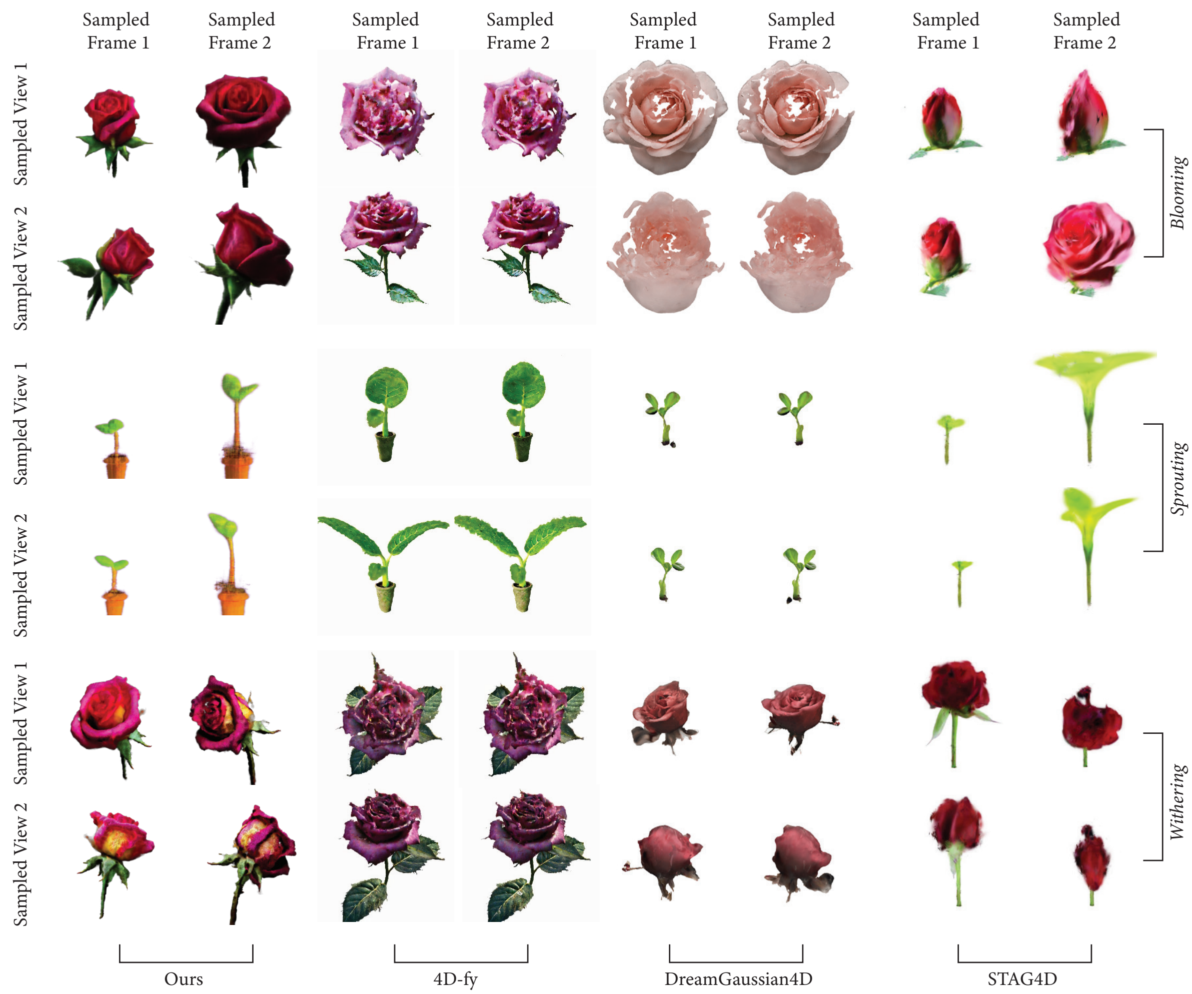

Figure 6. **Comparisons with 4D generation baselines.** We compare with several state-of-the-art 4D generation methods as they are the closest to our setting; however, none of these methods synthesize full dynamic object intrinsics. We observed that 4D-fy [2] and DreamGaussian4D [44] produce visually appealing static frames, but typically produce very small or unrealistic motions. STAG4D performs poorly in novel view renderings due to the limitation of the 3D reconstruction model.

state-of-the-art baselines.

6.1. Temporal Object Intrinsics Generation

We study 7 different natural phenomena of dynamic objects with significant intrinsics changes: "flower blooming", "flower withering", "plant sprouting", "candle burning", "icecream melting", "banana rotting", and "bread baking". Figure 5 shows part of the generation result of 4D object intrinsics from the proposed framework. More results can be found in the supplementary material.

Even for these complex dynamic phenomena with drastic object intrinsics transformation, our generation pipeline produces realistic synthesis results of the intrinsic properties with high temporal consistency in the dynamics motion.

6.2. Comparison with Prior Works

To the best of our knowledge, this work is the first to generate temporal object intrinsics for a natural process. Therefore, we choose baselines under the most similar task.

4D Generation Methods. We compare with recent 4D content generation methods for generic objects, on the 7 different natural phenomena described above. Note that the prior methods designed for this task cannot synthesize albedo or material properties as ours. Therefore, we only evaluate 4D renderings for comparison.

4D-fy [2] synthesizes the 4D content conditioned on text prompts, using guidance from video and image diffusion models and a multi-resolution hash encoding for representation. DreamGaussian4D [44] receives an image as input, and synthesizes 4D outputs with a 4D Gaussian Splatting representation. STAG4D [75] takes a video input and performs 4D reconstruction on the input video. We use the same reference video as ours to condition STAG4D. The detailed preparation of inputs for the baselines is described in the supplementary material.

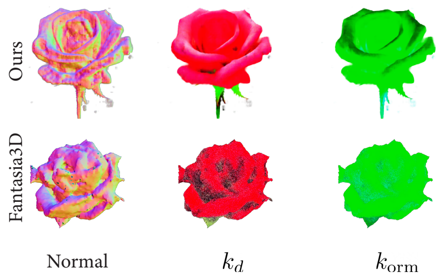

Figure 7. **Comparison with a static object intrinsics generation baseline.** We compare the state-of-the-art Fantasia3D [10], which synthesizes a “static” version of the temporal object intrinsics.

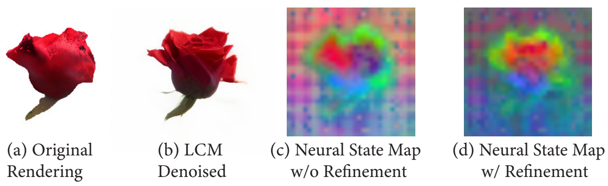

Figure 8. **Analyzing the influence of using consistency model to perform refinement.** Using the consistency model to do refinement (Sec. 4.2) helps mitigate the distribution gap between the raw RGB renderings and natural images and improves the quality of the output Neural State Maps (c-d).

Results are shown in Figure 6. Our method demonstrates higher rendering quality and better motion consistency compared to the baselines. 4D-fy fails to synthesize dynamic content in accordance with prompted motion concepts. DreamGaussian4D demonstrates severe temporal and static artifacts. STAG4D performs 4D reconstruction for a given input video and can fit the input view, but shows noticeable artifacts in novel views.

We perform a user study comparing the quality of our method with the baselines, where participants are asked to select the most preferred results based on three criteria: concept alignment, motion realism, and overall visual quality. The results of the average rate of preference among 95 participants are reported in Tab. 1, showing a clear preference for the proposed methods. Beyond user study, we report CLIP score [43] in Tab. 1 to reflect the objective visual quality and textual alignment. Additional results and detailed settings are included in the supplementary material.

3D Generation Methods with Material Modeling. Another line of work that tackles part of our task is 3D generation methods with decomposed geometry and appearance. In particular, we compare with Fantasia3D [10], which receives a text prompt as input (see the supplementary material for details). As shown in Figure 7, the baseline has a severe Janus problem in geometry and artifacts in texture, while our method synthesizes higher-quality materials.

6.3. Ablations and Validations

We ablate several important components in our pipeline.

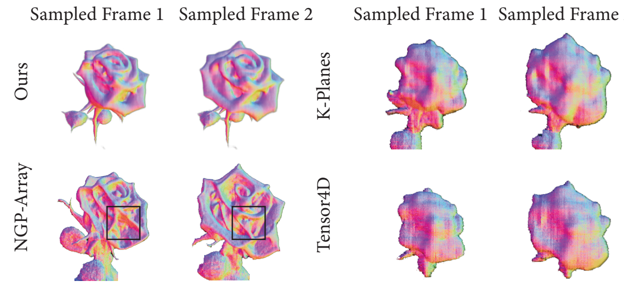

Figure 9. **Analysis of different 4D representations in guided generation.** We replace the hybrid 4D representation (Sec. 5.1) with three alternatives, resulting in either over-smoothing artifacts (K-Planes and Tensor4D) or temporal inconsistency (NGP-Array).

Denoising Renderings with Consistency Model. We study the effect of the operation to denoise with consistency model (Sec. 4.2) by visualizing the denoised RGB rendering and corresponding Neural State Maps in Figure 8. The original rendering of the dynamic mesh has a correct temporal state, yet its appearance does not fully align with the natural image distribution DINOv2 encoder F was trained on. By performing denoising, we can obtain a better Neural State Map depicting the temporal state more accurately.

Generation without Guidance. We show that using guidance during generation is necessary. Removing the Neural Template guidance in the pipeline leads to results in Figure 10, with visual artifacts of 3D inconsistency across views and the tiny motion across frames, suggesting the importance of the guidance.

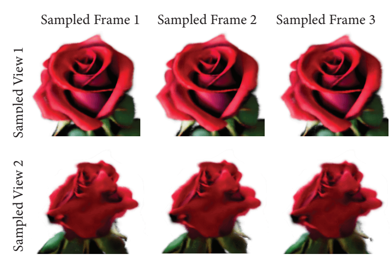

Figure 10. **Ablation of Neural Template guidance.**

Architecture of Hybrid 4D Representations. We further explore different variants of 4D representation used in the generation pipeline. We experiment with other possible representations, including pure K-Planes [13], Tensor4D [50], and an array of NGP encodings [37]. The results can be found at Figure 9. K-Planes and Tensor4D produce smooth outputs but struggle to represent the fine details. An array of NGP encodings is expressive enough but manifests poor temporal consistency, while the proposed representation achieves the best overall quality.

7. Conclusion

We have presented a novel and important task of generating 4D temporal object intrinsics only from the guidance of 2D foundation models, proposed a new representation of Neural Template, and explored its use in generating temporal object intrinsics for different phenomena.

Acknowledgments. The paper's title is inspired by the artwork, *Nacer y Morir de una Rosa*, by Colombian artist Rosa Navarro, recently on exhibit at the Orange County Museum of Art in 2023. We thank Haian Jin, Zizhang Li, and Ruocheng Wang for insightful discussions and feedback on the paper. This work is in part supported by NSF RI #2211258, #2338203, ONR MURI N00014-22-1-2740, and Samsung. YZ is in part supported by the Stanford Interdisciplinary Graduate Fellowship.

References

[1] Mohammadreza Armandpour, Huangjie Zheng, Ali Sadeghian, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*, 2023. 2

[2] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *arXiv preprint arXiv:2311.17984*, 2023. 2, 3, 6, 7

[3] Jonathan T Barron and Jitendra Malik. Color constancy, intrinsic images, and shape estimation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pages 57–70. Springer, 2012. 2

[4] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE TPAMI*, 37(8):1670–1687, 2014. 2

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[6] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *ICCV*, pages 12684–12694, 2021. 2

[7] Brent Burley. Physically-based shading at disney. In *ACM TOG*, 2012. 5

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2

[9] Ehtzaz Chaudhry, Algirdas Noreika, Lihua You, Jian-Jun Zhang, Jian Chang, Hassan Ugail, Alexander Malyshev, Alfonso Carriazo, Andrés Iglesias, Zulfiqar Habib, Allah Bux Sargano, and Habibollah Haron. Modelling and simulation of lily flowers using pde surfaces. *International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, pages 1–8, 2019. 1, 2

[10] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*, 2023. 2, 8

[11] Kangle Deng, Timothy Omernick, Alexander Weiss, Deva Ramanan, Jun-Yan Zhu, Tinghui Zhou, and Maneesh Agrawala. Flashtex: Fast relightable mesh texturing with lightcontrolnet. *arXiv preprint arXiv:2402.13251*, 2024. 2

[12] Wenqi Dong, Bangbang Yang, Lin Ma, Xiao Liu, Liyuan Cui, Hujun Bao, Yuewen Ma, and Zhaopeng Cui. Coin3d: Controllable and interactive 3d assets generation with proxy-guided conditioning. *arXiv preprint arXiv:2405.08054*, 2024. 2

[13] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023. 5, 8

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 2

[15] Quankai Gao, Qiangeng Xu, Zhe Cao, Ben Mildenhall, Wenchao Ma, Le Chen, Danhang Tang, and Ulrich Neumann. Gaussianflow: Splatting gaussian dynamics for 4d content creation. *arXiv preprint arXiv:2403.12365*, 2024. 2

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 4

[17] Takeo Igarashi, Tomer Moscovich, and John F Hughes. As-rigid-as-possible shape manipulation. *ACM transactions on Graphics (TOG)*, 24(3):1134–1141, 2005. 4

[18] Takashi Ijiri, Shigeru Owada, Makoto Okabe, and Takeo Igarashi. Floral diagrams and inflorescences: interactive flower modeling using botanical structural constraints. In *ACM SIGGRAPH 2007 Courses*, 2007. 1, 2

[19] Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360° dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023. 2

[20] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. Tensoir: Tensorial inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2023. 2

[21] Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion. *ArXiv*, abs/2406.07520, 2024. 2

[22] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 5

[23] Ladislav Kavan and Jiří Žára. Spherical blend skinning: a real-time deformation of articulated models. In *Proceedings of the 2005 symposium on Interactive 3D graphics and games*, pages 9–16, 2005. 3

[24] Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Skinning with dual quaternions. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games*, pages 39–46, 2007. 3

[25] Peter Kocsis, Julien Philip, Kalyan Sunkavalli, Matthias Nießner, and Yannick Hold-Geoffroy. Lightit: Illumination modeling and control for diffusion models. In *CVPR*, 2024. 2

[26] Hsin-Ying Lee, Hung-Yu Tseng, Hsin-Ying Lee, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable dense prediction. In *CVPR*, 2024. 2

[27] Zhiqi Li, Yiming Chen, and Peidong Liu. Dreammesh4d: Video-to-4d generation with sparse-controlled gaussian-mesh hybrid representation. *arXiv preprint arXiv:2410.06756*, 2024. 2

[28] Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, and Jiajun Wu. Learning the 3d fauna of the web. *arXiv preprint arXiv:2401.02400*, 2024. 2

[29] Hanwen Liang, Yuyang Yin, Dejia Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. *arXiv preprint arXiv:2405.16645*, 2024. 2

[30] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8576–8588, 2024. 2

[31] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, pages 9298–9309, 2023. 2, 4

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 2, 3

[33] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023. 5

[34] Ishit Mehta, Manmohan Chandraker, and Ravi Ramamoorthi. A level set theory for neural implicit evolution under explicit flows. *arXiv preprint arXiv:2204.07159*, 2022. 5

[35] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 2

[36] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 5

[37] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 41(4):1–15, 2022. 5, 8

[38] Tiago Novello, Vinícius da Silva, Guilherme Schardong, Luiz Schirmer, Hélio Lopes, and Luiz Velho. Neural implicit surface evolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14279–14289, 2023. 5

[39] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 3, 5

[40] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae, 2021. 2

[41] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 1, 2, 3

[42] Przemyslaw Prusinkiewicz and Aristid Lindenmayer. *The Algorithmic Beauty of Plants*. Springer-Verlag, 1990. 1, 2

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 2021. 8

[44] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 2, 3, 6, 7

[45] Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, et al. L4gm: Large 4d gaussian reconstruction model. *arXiv preprint arXiv:2406.10324*, 2024. 2

[46] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 2

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 5

[48] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2

[49] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 6

[50] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16632–16642, 2023. 8

[51] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12469–12480, 2023. 4

[52] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. 5

[53] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, pages 109–116, 2007. 4

[54] Keqiang Sun, Dor Litvak, Yunzhi Zhang, Hongsheng Li, Jiajun Wu, and Shangzhe Wu. Ponymation: Learning 3d animal motions from unlabeled online videos. *arXiv preprint arXiv:2312.13604*, 2023. 2

[55] Qi Sun, Zhiyang Guo, Ziyu Wan, Jing Nathan Yan, Shengming Yin, Wengang Zhou, Jing Liao, and Houqiang Li. Eg4d: Explicit generation of 4d object without score distillation. *arXiv preprint arXiv:2405.18132*, 2024. 2

[56] Lukas Uzolas, Elmar Eisemann, and Petr Kellnhofer. Motiondreamer: Zero-shot 3d mesh animation from video diffusion models. *arXiv preprint arXiv:2405.20155*, 2024. 2

[57] Nicolás Violante, Alban Gauthier, Stavros Diolatzis, Thomas Leimkühler, and George Drettakis. Physically-based lighting for 3d generative models of cars. In *Computer Graphics Forum*, page e15011. Wiley Online Library, 2024. 2

[58] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 3

[59] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 4

[60] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 5

[61] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3295–3306, 2023. 5

[62] Yikai Wang, Xinzhou Wang, Zilong Chen, Zhengyi Wang, Fuchun Sun, and Jun Zhu. Vidu4d: Single generated video to high-fidelity 4d reconstruction with dynamic gaussian surfels. *arXiv preprint arXiv:2405.16822*, 2024. 2

[63] Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Dove: Learning deformable 3d objects by watching videos. *International Journal of Computer Vision*, 131(10):2623–2634, 2023. 4

[64] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *CVPR*, pages 8792–8802, 2023. 2

[65] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024. 2

[66] Xudong Xu, Zhaoyang Lyu, Xingang Pan, and Bo Dai. Matlaber: Material-aware text-to-3d via latent brdf auto-encoder. *arXiv preprint arXiv:2308.09278*, 2023. 2

[67] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 2

[68] Zeyu Yang, Zijie Pan, Chun Gu, and Li Zhang. Diffusion^2: Dynamic 3d content generation via score composition of orthogonal diffusion models. *arXiv preprint arXiv:2404.02148*, 2024. 2

[69] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 4, 6

[70] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Lassie: Learning articulated shapes from sparse image ensemble via 3d part discovery. *Advances in Neural Information Processing Systems*, 35:15296–15308, 2022. 2

[71] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Hi-lassie: High-fidelity articulated shape and skeleton discovery from sparse image ensemble. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4853–4862, 2023. 2

[72] Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Laszlo A Jeni, Sergey Tulyakov, and Hsin-Ying Lee. 4real: Towards photorealistic 4d scene generation via video diffusion models. *arXiv preprint arXiv:2406.07472*, 2024. 2

[73] Yu-Jie Yuan, Leif Kobbelt, Jiwen Liu, Yuan Zhang, Pengfei Wan, Yu-Kun Lai, and Lin Gao. 4dynamic: Text-to-4d generation with hybrid priors. *arXiv preprint arXiv:2407.12684*, 2024. 2

[74] Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. Dilightnet: Fine-grained lighting control for diffusion-based image generation. *ArXiv*, abs/2402.11929, 2024. 2

[75] Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. Stag4d: Spatial-temporal anchored generative 4d gaussians. In *European Conference on Computer Vision*, pages 163–179. Springer, 2024. 6, 7

[76] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. PhySG: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *CVPR*, 2021. 2

[77] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2, 3, 4, 6

[78] Xiuming Zhang, Pratul P. Srinivasan, Boyang Deng, Paul Debevec, William T. Freeman, and Jonathan T. Barron. Nerfactor: neural factorization of shape and reflectance under an unknown illumination. *ACM TOG*, 40(6):1–18, 2021. 2

[79] Yunzhi Zhang, Shangzhe Wu, Noah Snavely, and Jiajun Wu. Seeing a rose in five thousand ways. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 962–971, 2023. 2

[80] Yuqing Zhang, Yuan Liu, Zhiyu Xie, Lei Yang, Zhongyuan Liu, Mengzhou Yang, Runze Zhang, Qilong Kou, Cheng Lin, Wenping Wang, et al. Dreammat: High-quality pbr material generation with geometry-and light-aware diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024. 2

[81] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*, 2023. 2