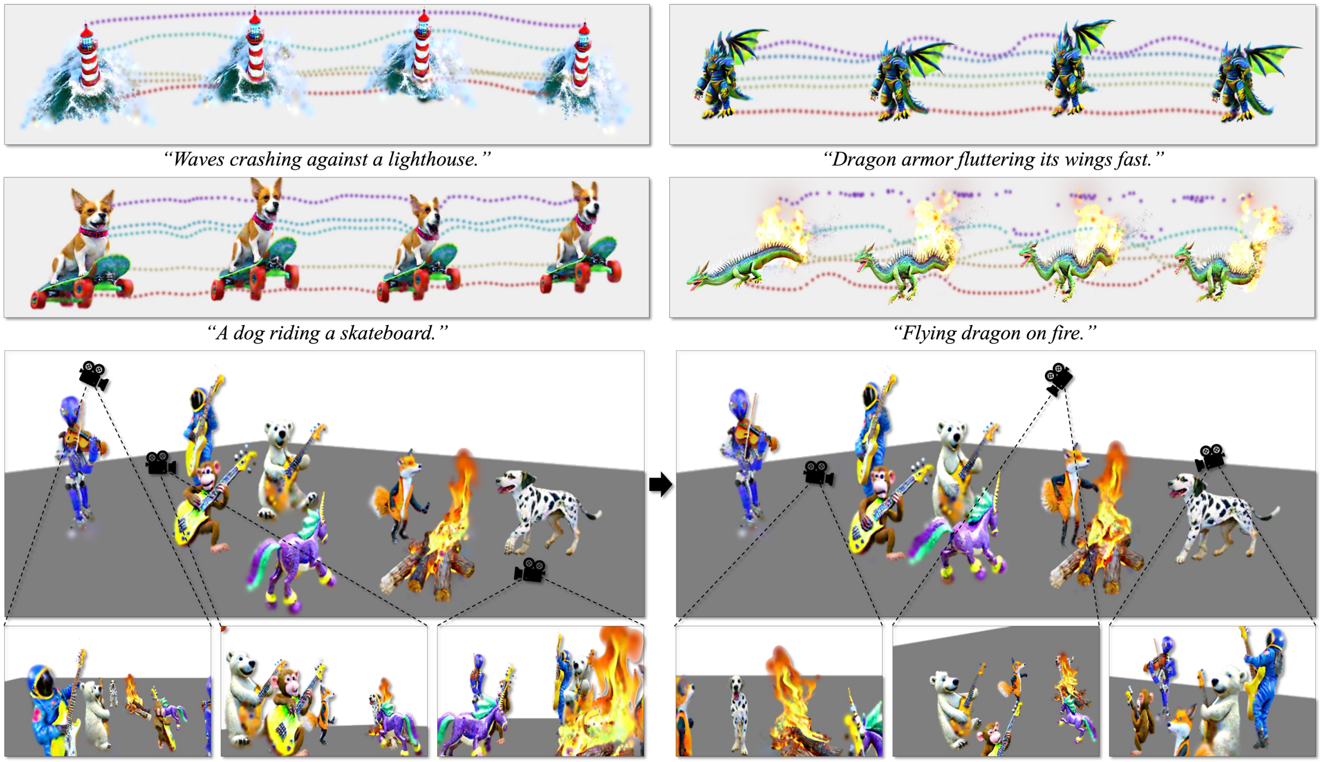
Align Your Gaussians: Text-to-4D with Dynamic 3D Gaussians and Composed Diffusion Models

Huan Ling[1,2,3] * Seung Wook Kim[1,2,3] * Antonio Torralba[4] Sanja Fidler[1,2,3] Karsten Kreis[1]

[1]NVIDIA [2]Vector Institute [3]University of Toronto [4]MIT

Project page: https://research.nvidia.com/labs/toronto-ai/AlignYourGaussians/



"Waves crashing against a lighthouse."

"Dragon armor fluttering its wings fast."

"A dog riding a skateboard."

"Flying dragon on fire."

Figure 1. **Text-to-4D synthesis with *Align Your Gaussians (AYG)*.** *Top:* Different dynamic 4D sequences. Dotted lines represent dynamics of deformation field. *Bottom:* Multiple dynamic 4D objects are composed within a large dynamic scene; two time frames shown.

Abstract

*Text-guided diffusion models have revolutionized image and video generation and have also been successfully used for optimization-based 3D object synthesis. Here, we instead focus on the underexplored text-to-4D setting and synthesize dynamic, animated 3D objects using score distillation methods with an additional temporal dimension. Compared to previous work, we pursue a novel compositional generation-based approach, and combine text-to-image, text-to-video, and 3D-aware multiview diffusion models to provide feedback during 4D object optimization, thereby simultaneously enforcing temporal consistency, high-quality visual appearance and realistic geometry. Our method, called **Align Your Gaussians (AYG)**, leverages dynamic 3D Gaussian Splatting with deformation fields as 4D representation. Crucial to AYG is a novel method to regularize the distribution of the moving 3D Gaussians and thereby stabilize the optimization and induce motion. We also propose a motion amplification mechanism as well as a new autoregressive synthesis scheme to generate and combine multiple 4D sequences for longer generation. These techniques allow us to synthesize vivid dynamic scenes, outperform previous work qualitatively and quantitatively and achieve state-of-the-art text-to-4D performance. Due to the Gaussian 4D representation, different 4D animations can be seamlessly combined, as we demonstrate. AYG opens up promising avenues for animation, simulation and digital content creation as well as synthetic data generation.*

1. Introduction

Generative modeling of dynamic 3D scenes has the potential to revolutionize how we create games, movies, simu-

[*] Equal contribution.

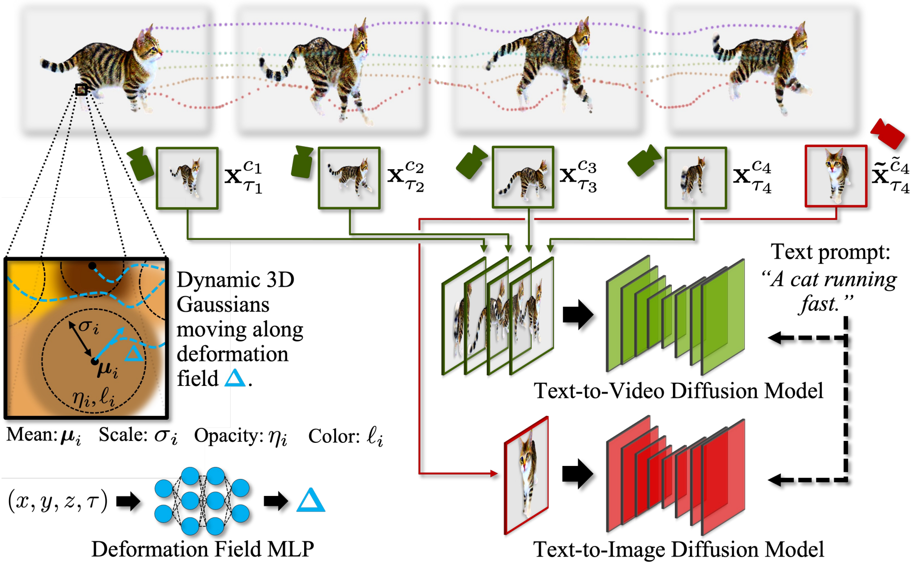

$\mathbf{x}_{\tau_1}^{c_1}$
$\mathbf{x}_{\tau_2}^{c_2}$
$\mathbf{x}_{\tau_3}^{c_3}$
$\mathbf{x}_{\tau_4}^{c_4}$
$\tilde{\mathbf{x}}_{\tau_4}^{\tilde{c}_4}$
σ_i
$\boldsymbol{\mu}_i$
η_i, ℓ_i
Dynamic 3D Gaussians moving along deformation field △.
Mean: $\boldsymbol{\mu}_i$ Scale: σ_i Opacity: η_i Color: ℓ_i
Text prompt: *"A cat running fast."*
Text-to-Video Diffusion Model
Text-to-Image Diffusion Model
(x, y, z, τ)
△
Deformation Field MLP

Figure 2. **Text-to-4D synthesis with AYG.** We generate dynamic 4D scenes via score distillation. We initialize the 4D sequence from a static 3D scene (generated first, Fig. 3), which is represented by 3D Gaussians with means $\boldsymbol{\mu}_i$, scales σ_i, opacities η_i and colors ℓ_i. Consecutive rendered frames $\mathbf{x}_{\tau_j}^{c_j}$ from the 4D sequence at times τ_j and camera positions c_j are diffused and fed to a text-to-video diffusion model [7] (**green arrows**), which provides a distillation gradient that is backpropagated through the rendering process into a deformation field $\triangle(x, y, z, \tau)$ (dotted lines) that captures scene motion. Simultaneously, random frames $\tilde{\mathbf{x}}_{\tau_j}^{\tilde{c}_j}$ are diffused and given to a text-to-image diffusion model [70] (**red arrows**) whose gradients ensure that high visual quality is maintained frame-wise.

lations, animations and entire virtual worlds. Many works have shown how a wide variety of 3D objects can be synthesized via score distillation techniques [10, 11, 31, 41, 52, 62, 79, 85, 88, 92, 109], but they typically only synthesize static 3D scenes, although we live in a moving, dynamic world. While image diffusion models have been successfully extended to video generation [1, 7, 22, 28, 78, 90, 91, 107], there is little research on similarly extending 3D synthesis to 4D generation with an additional temporal dimension.

We propose *Align Your Gaussians (AYG)*, a novel method for 4D content creation. In contrast to previous work [79], we leverage dynamic 3D Gaussians [36] as backbone 4D representation, where a deformation field [59, 63] captures scene dynamics and transforms the collection of 3D Gaussians to represent object motion. AYG takes a compositional generation-based perspective and leverages the combined gradients of latent text-to-image [70], text-to-video [7] and 3D-aware text-to-multiview-image [76] diffusion models in a score distillation-based synthesis framework. A 3D-aware multiview diffusion model and a regular text-to-image model are used to generate an initial high-quality 3D shape. Afterwards, we compose the gradients of a text-to-video and a text-to-image model; the gradients of the text-to-video model optimize the deformation field to capture temporal dynamics, while the text-to-image model ensures that high visual quality is maintained for all time frames (Fig. 2). To this end, we trained a dedicated text-to-video model; it is conditioned on the frame rate and can create useful gradients both for short and long time intervals, which allows us to generate long and smooth 4D sequences.

We developed several techniques to ensure stable optimization and learn vivid dynamic 4D scenes in AYG: We employ a novel regularization method that uses a modified version of the Jensen-Shannon divergence to regularize the locations of the 3D Gaussians such that the mean and variance of the set of 3D Gaussians is preserved as they move. Furthermore, we use a motion amplification method that carefully scales the gradients from the text-to-video model and enhances motion. To extend the length of the 4D sequences or combine different dynamic scenes with changing text guidance, we introduce an autoregressive generation scheme which interpolates the deformation fields of consecutive sequences. We also propose a new view-guidance method to generate consistent 3D scenes for initialization of the 4D stage, and we leverage the concurrent classifier score distillation method [102].

We find that AYG can generate diverse, vivid, detailed and 3D-consistent dynamic scenes (Fig. 1), achieving state-of-the-art text-to-4D performance. We also show long, autoregressively extended 4D scenes, including ones with varying text guidance, which has not been demonstrated before. A crucial advantage of AYG's 4D Gaussian backbone representation is that different 4D animations can trivially be combined and composed together, which we also show.

We envision broad applications in digital content creation, where AYG takes a step beyond the literature on text-to-3D and captures our world's rich dynamics. Moreover, AYG can generate 4D scenes with exact tracking labels for free, a promising feature for synthetic data generation.

Contributions. *(i)* We propose AYG, a system for text-to-4D content creation leveraging dynamic 3D Gaussians with deformation fields as 4D representation. *(ii)* We show how to tackle the text-to-4D task through score distillation within a new compositional generation framework, combining 2D, 3D, and video diffusion models. *(iii)* To scale AYG, we introduce a novel regularization method and a new motion amplification technique. *(iv)* Experimentally, we achieve state-of-the-art text-to-4D performance and generate high-quality, diverse, and dynamic 4D scenes. *(v)* For the first time, we also show how our 4D sequences can be extended in time with a new autoregressive generation scheme and even creatively composed in large scenes.

2. Background

3D Gaussian Splatting [36] represents 3D scenes by N 3D Gaussians with positions $\boldsymbol{\mu}_i$, covariances Σ_i, opacities η_i and colors ℓ_i (Fig. 2). Rendering corresponds to projection of the 3D Gaussians onto the 2D camera's image plane, producing 2D Gaussians with projected means $\hat{\boldsymbol{\mu}}_i$ and covariances $\hat{\Sigma}_i$. The color $\mathcal{C}(\mathbf{p})$ of image pixel $\mathbf{p}$ can be calculated through point-based volume rendering [111] as

$$\mathcal{C}(\mathbf{p}) = \sum_{i=1}^{N} \ell_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{1}$$

$$\alpha_i = \eta_i \exp\left[-\frac{1}{2}(\mathbf{p} - \hat{\boldsymbol{\mu}}_i)^\top \hat{\Sigma}_i^{-1} (\mathbf{p} - \hat{\boldsymbol{\mu}}_i)\right], \tag{2}$$

where j iterates over the Gaussians along the ray through the scene from pixel $\mathbf{p}$ until Gaussian i. To accelerate rendering, the image plane can be divided into tiles, which are processed in parallel. Initially proposed for 3D scene reconstruction, 3D Gaussian Splatting uses gradient-based thresholding to densify areas that need more Gaussians to capture fine details, and unnecessary Gaussians with low opacity are pruned every few thousand optimization steps.

Diffusion Models and Score Distillation Sampling. Diffusion-based generative models (DMs) [18, 27, 57, 80, 81] use a forward diffusion process that gradually perturbs data, such as images or entire videos, towards entirely random noise, while a neural network is learnt to denoise and reconstruct the data. DMs have also been widely used for score distillation-based generation of 3D objects [62]. In that case, a 3D object, represented for instance by a neural radiance field (NeRF) [54] or 3D Gaussians [36], like here, with parameters $\boldsymbol{\theta}$ is rendered from different camera views and the renderings $\mathbf{x}$ are diffused and given to a text-to-image DM. In the score distillation sampling (SDS) framework, the DM's denoiser is then used to construct a gradient that is backpropagated through the differentiable rendering process g into the 3D scene representation and updates the scene representation to make the scene rendering look more realistic, like images modeled by the DM. Rendering and using DM feedback from many different camera perspectives then encourages the scene representation to form a geometrically consistent 3D scene. The SDS gradient [62] is

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{SDS}}(\mathbf{x} = g(\boldsymbol{\theta})) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[w(t)\left(\hat{\boldsymbol{\epsilon}}_\phi(\mathbf{z}_t, v, t) - \boldsymbol{\epsilon}\right)\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}\right],$$

where $\mathbf{x}$ denotes the 2D rendering, t is the time up to which the diffusion is run to perturb $\mathbf{x}$, $w(t)$ is a weighting function, and $\mathbf{z}_t$ is the perturbed rendering. Further, $\hat{\boldsymbol{\epsilon}}_\phi(\mathbf{z}_t, v, t)$ is the DM's denoiser neural network that predicts the diffusion noise $\boldsymbol{\epsilon}$. It is conditioned on $\mathbf{z}_t$, the diffusion time t and a text prompt v for guidance. Classifier-free guidance (CFG) [26] typically amplifies the text conditioning.

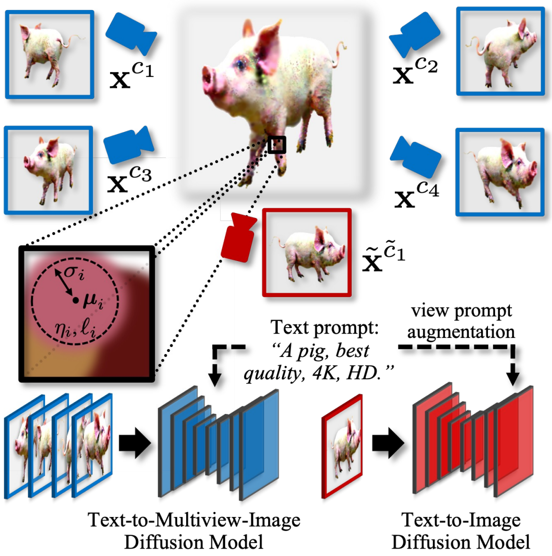

$\mathbf{x}^{c_1}$
$\mathbf{x}^{c_2}$
$\mathbf{x}^{c_3}$
$\mathbf{x}^{c_4}$
σ_i
$\boldsymbol{\mu}_i$
η_i, ℓ_i
$\tilde{\mathbf{x}}^{\tilde{c}_1}$
Text prompt: "A pig, best quality, 4K, HD."
view prompt augmentation
Text-to-Multiview-Image Diffusion Model
Text-to-Image Diffusion Model

Figure 3. In **AYG's initial 3D stage** we synthesize a static 3D scene leveraging a text-guided multiview diffusion model [76] and a regular text-to-image model [70]. The text-to-image model receives viewing angle-dependent text prompts and leverages view guidance (Sec. 3.4). See Fig. 2 for 4D stage and descriptions.

2.1. Related Work

See Supp. Material for an extended discussion. Here, we only briefly mention the most relevant related literature.

As discussed, AYG builds on text-driven image [5, 14, 21, 61, 67, 70, 72, 98], video [1, 7, 22, 25, 28, 38, 78, 90, 91, 94, 107] and 3D-aware DMs [42, 44, 45, 56, 64, 75, 76, 104], uses score distillation sampling [10, 11, 17, 31, 41, 48, 52, 62, 85, 88, 92, 96, 109] and leverages 3D Gaussian Splatting [36] as well as deformation fields [8, 59, 60, 63, 84] for its 4D representation. The concurrent works DreamGaussian [83], GSGEN [12] and GaussianDreamer [101] use 3D Gaussian Splatting to synthesize static 3D scenes, but do not consider dynamics. Dynamic 3D Gaussian Splatting has been used for 4D reconstruction [50, 93, 110], but not for 4D generation. The idea to compose the gradients of multiple DMs has been used before for controllable image generation [19, 43], but has received little attention in 3D or 4D synthesis.

Most related to AYG is *Make-A-Video3D (MAV3D)* [79], to the best of our knowledge the only previous work that generates dynamic 4D scenes with score distillation. MAV3D uses NeRFs with HexPlane [9] features as 4D representation, in contrast to AYG's dynamic 3D Gaussians, and it does not disentangle its 4D representation into a static 3D representation and a deformation field modeling dynamics. MAV3D's representation prevents it from composing multiple 4D objects into large dynamic scenes, which our 3D Gaussian plus deformation field representation easily enables, as we show. Moreover, MAV3D's sequences are limited in time, while we show a novel autoregressive generation scheme to extend our 4D sequences.

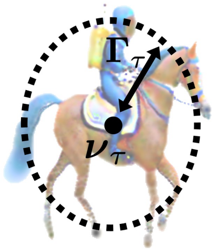


Figure 4. **AYG's JSD-based regularization** of the evolving 4D Gaussians (see Sec. 3.4) calculates the 3D mean $\boldsymbol{\nu}_\tau$ and diagonal covariance matrix $\boldsymbol{\Gamma}_\tau$ of the set of dynamic 3D Gaussians at different times τ of the 4D sequence and regularizes them to not vary too much.

AYG outperforms MAV3D qualitatively and quantitatively and synthesizes significantly higher-quality 4D scenes. Our novel compositional generation-based approach contributes to this, which MAV3D does not pursue. Finally, instead of regular SDS, used by MAV3D, in practice AYG employs classifier score distillation [102] (see Sec. 3.4).

3. Align Your Gaussians

In Sec. 3.1, we present AYG's 4D representation, and in Sec. 3.2, we introduce its compositional generation framework with multiple DMs. In Sec. 3.3, we lay out AYG's score distillation framework in practice, and in Sec. 3.4, we discuss several novel methods and extensions to scale AYG.

3.1. AYG's 4D Representation

AYG's 4D representation combines 3D Gaussian Splatting [36] with deformation fields [59, 63] to capture the 3D scene and its temporal dynamics in a disentangled manner. Specifically, each 4D scene consists of a set of N 3D Gaussians as in Sec. 2. Following Kerbl et al. [36], we also use two degrees of spherical harmonics to model view-dependent effects, this is, directional color, and thereby improve the 3D Gaussians' expressivity. Moreover, we restrict the 3D Gaussians' covariance matrices to be isotropic with scales σ_i. We made this choice as our 3D Gaussians move as a function of time and learning expressive dynamics is easier for spherical Gaussians. We denote the collection of learnable parameters of our 3D Gaussians as $\boldsymbol{\theta}$. The scene dynamics are modeled by a deformation field $\boldsymbol{\Delta}_{\boldsymbol{\Phi}}(x, y, z, \tau) = (\Delta x, \Delta y, \Delta z)$, defined through a multi-layer perceptron (MLP) with parameters $\boldsymbol{\Phi}$. Specifically, for any 3D location (x, y, z) and time τ, the deformation field predicts a displacement $(\Delta x, \Delta y, \Delta z)$. The 3D Gaussians smoothly follow these displacements to represent a moving and deforming 4D scene (Fig. 2). Note that in practice we preserve the initial 3D Gaussians for the first frame, *i.e.* $\boldsymbol{\Delta}_{\boldsymbol{\Phi}}(x, y, z, 0) = (0, 0, 0)$, by setting $\boldsymbol{\Delta}_{\boldsymbol{\Phi}}(x, y, z, \tau) = (\xi(\tau)\Delta x, \xi(\tau)\Delta y, \xi(\tau)\Delta z)$ where $\xi(\tau) = \tau^{0.35}$ such that $\xi(0) = 0$ and $\xi(1) = 1$. Following Luiten et al. [50], we regularize the deformation field so that nearby Gaussians deform similarly ("rigidity regularization", see Supp. Mat.).

Apart from the intuitive decomposition into a backbone 3D representation and a deformation field to model dynamics, a crucial advantage of AYG's dynamic 3D Gaussian-based representation is that different dynamic scenes, each with its own set of Gaussians and deformation field, can be easily combined, thereby enabling promising 3D dynamic content creation applications (see Fig. 1). This is due to the explicit nature of this representation, in contrast to typical NeRF-based representations. Moreover, learning 4D scenes with score distillation requires many scene renderings. This also makes 3D Gaussians ideal due to their rendering efficiency [36]. Note that early on we also explored MAV3D's HexPlane- and NeRF-based 4D representation [79], but we were not able to achieve satisfactory results.

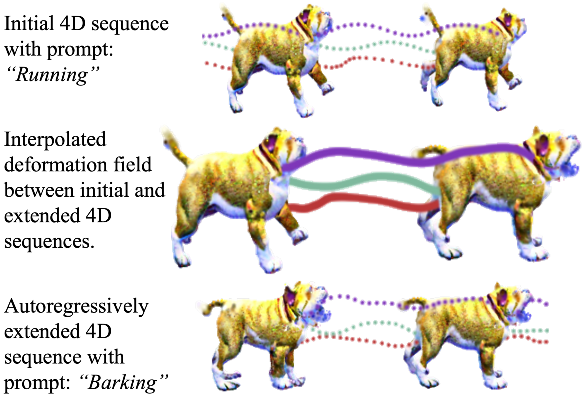
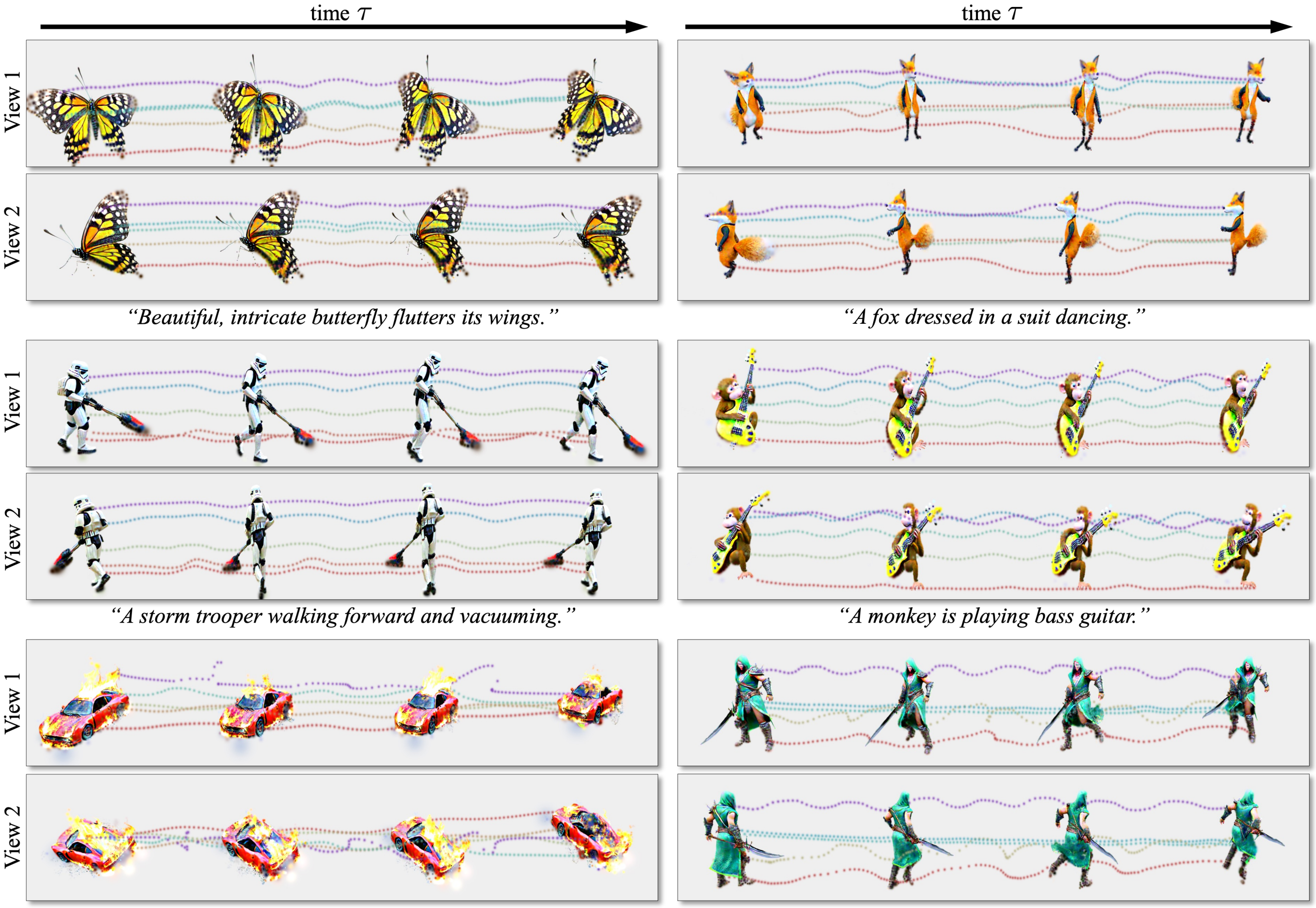

Initial 4D sequence with prompt: *"Running"*
Interpolated deformation field between initial and extended 4D sequences.
Autoregressively extended 4D sequence with prompt: *"Barking"*

Figure 5. **AYG's autoregressive extension scheme** interpolates the deformation fields of an initial and an extended 4D sequence within an overlap interval between the two sequences (Sec. 3.4).

3.2. Text-to-4D as Compositional Generation

We would like AYG's synthesized dynamic 4D scenes to be of high visual quality, be 3D-consistent and geometrically correct, and also feature expressive and realistic temporal dynamics. This suggests to compose different text-driven DMs during the distillation-based generation to capture these different aspects. *(i)* We use the text-to-image model Stable Diffusion (SD) [70], which has been trained on a broad set of imagery and provides a strong general image prior. *(ii)* We also utilize the 3D-aware text-conditioned multi-view DM MVDream [76], which generates multi-view images of 3D objects, was fine-tuned from SD on the object-centric 3D dataset Objaverse [15, 16] and provides a strong 3D prior. It defines a distribution over four multiview-consistent images corresponding to object renderings from four different camera perspectives $c_1, ..., c_4$. Moreover, we train a text-to-video DM, following VideoLDM [7], but with a larger text-video dataset (HDVG-130M [90] and Webvid-10M [4]) and additional conditioning on the videos' frame rate (see Supp. Material for details). This video DM provides temporal feedback when rendering 2D frame sequences from our dynamic 4D scenes. All used DMs are *latent* DMs [70, 86], which means that in practice we first encode renderings of our 4D scenes into the models' latent spaces, calculate score distillation gradients there, and backpropagate them through the models' encoders. All DMs leverage the SD 2.1 backbone and share the same encoder. To keep the notation simple, we do



"Beautiful, intricate butterfly flutters its wings."

"A fox dressed in a suit dancing."

"A storm trooper walking forward and vacuuming."

"A monkey is playing bass guitar."

"Burning sports car."

"Assassin with sword running fast, portrait, game, unreal, 4K, HD."

Figure 6. **Text-to-4D synthesis with AYG.** Various samples shown in two views each. Dotted lines denote deformation field dynamics.

not explicitly incorporate the encoding into our mathematical description below and the visualizations (Figs. 2 and 3).

We disentangle optimization into first synthesizing a static 3D Gaussian-based object $\boldsymbol{\theta}$, and then learning the deformation field $\boldsymbol{\Phi}$ to add scene dynamics.

Stage 1: 3D Synthesis (Fig. 3). We first use MVDream's multiview image prior to generate a static 3D scene via score distillation (Supp. Mat. for details). Since MVDream on its own would generate objects in random orientations, we enforce a canonical pose by combining MVDream's gradients with those of regular SD, while augmenting the text-conditioning for SD with directional texts "front view", "side view", "back view" and "overhead view" [62]. Formally, we can derive a score distillation gradient (see Sec. 3.3) by minimizing the reverse Kulback-Leibler divergence (KLD) from the rendering distribution to the product of the composed MVDream and SD model distributions

$$\mathrm{KL}\left(q_{\boldsymbol{\theta}}\left(\{\mathbf{z}^{c_i}\}_4, \{\tilde{\mathbf{z}}^{\tilde{c}_j}\}_K\right) \middle|\middle| p^{\alpha}_{\mathrm{3D}}\left(\{\mathbf{z}^{c_i}\}_4\right)\prod_{j=1}^{K} p^{\beta}_{\mathrm{im}}\left(\tilde{\mathbf{z}}^{\tilde{c}_j}\right)\right),$$

similar to Poole et al. [62] (App. A.4). Here, $p_{\mathrm{3D}}(\{\mathbf{z}^{c_i}\}_4)$ represents the MVDream-defined multiview image distribution over four diffused renderings from camera views c_i, denoted as the set $\{\mathbf{z}^{c_i}\}_4$ (we omit the diffusion time t subscript for brevity). Moreover, $p_{\mathrm{im}}(\tilde{\mathbf{z}}^{\tilde{c}_j})$ is the SD-based general image prior and $\{\tilde{\mathbf{z}}^{\tilde{c}_j}\}_K$ is another set of K diffused scene renderings. In principle, the renderings for SD and MVDream can be from different camera angles c_i and $\tilde{c}_j$, but in practice we choose $K{=}4$ and use the same renderings. Furthermore, α and β are adjustable temperatures of the distributions p_{3D} and p_{im}, and $q_{\boldsymbol{\theta}}$ denotes the distribution over diffused renderings defined by the underlying 3D scene representation $\boldsymbol{\theta}$, which is optimized through the differentiable rendering. We also use the Gaussian densification method discussed in Sec. 2 (see Supp. Material).

Stage 2: Adding Dynamics for 4D Synthesis (Fig. 2). While in stage 1, we only optimize the 3D Gaussians, in stage 2, the main 4D stage, we optimize (only) the deformation field $\boldsymbol{\Phi}$ to capture motion and extend the static 3D scene to a dynamic 4D scene with temporal dimension τ. To this end, we compose the text-to-image and text-to-video DMs and formally minimize a reverse KLD of the form

$$\mathrm{KL}\left(q_{\boldsymbol{\Phi}}\left(\{\mathbf{z}^{c_i}_{\tau_i}\}_F, \{\tilde{\mathbf{z}}^{\tilde{c}_j}_{\tilde{\tau}_j}\}_M\right) \middle|\middle| p^{\gamma}_{\mathrm{vid}}\left(\{\mathbf{z}^{c_i}_{\tau_i}\}_F\right)\prod_{j=1}^{M} p^{\kappa}_{\mathrm{im}}\left(\tilde{\mathbf{z}}^{\tilde{c}_j}_{\tilde{\tau}_j}\right)\right),$$

where $p_{\mathrm{vid}}(\{\mathbf{z}^{c_i}_{\tau_i}\}_F)$ is the video DM-defined distribution

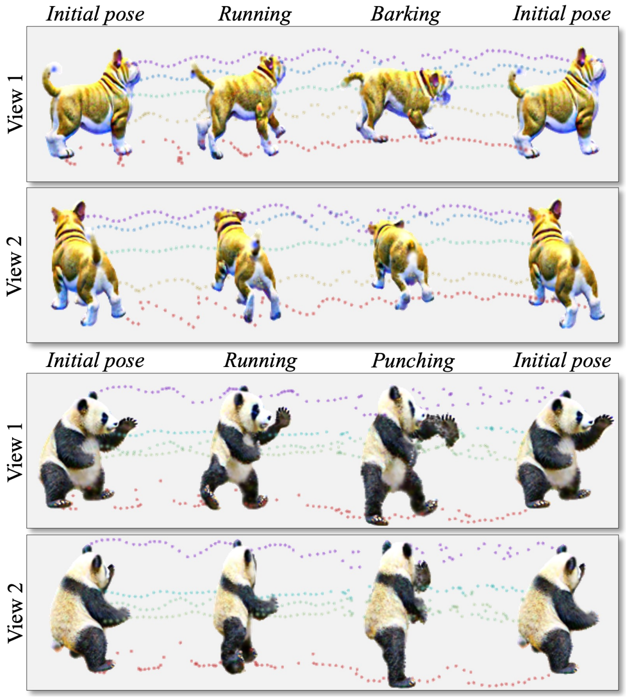


Figure 7. **Autoregressively extended text-to-4D synthesis.** AYG is able to autoregressively extend dynamic 4D sequences, combine sequences with different text-guidance, and create looping animations, returning to the initial pose (also see Supp. Video).

over F 4D scene renderings $\{\mathbf{z}^{c_i}_{\tau_i}\}_F$ taken at times τ_i and camera angles c_i (F=16 for our model). Similar to before, M additional renderings are given to the SD-based general image prior, and γ and κ are temperatures. The renderings $\{\tilde{\mathbf{z}}^{\tilde{c}_j}_{\tilde{\tau}_j}\}_M$ fed to regular SD can be taken at different times $\tilde{\tau}_j$ and cameras $\tilde{c}_j$ than the video model frames, but in practice M=4 and we use three random renderings as well as the 8th middle frame among the ones given to the video model. $q_{\mathbf{\Phi}}$ defines the distribution over renderings by the 4D scene with the learnable deformation field parameters $\mathbf{\Phi}$. We could render videos from the 4D scene with a fixed camera, but in practice dynamic cameras, *i.e.* varying c_i, help to learn more vivid 4D scenes, similar to Singer et al. [79].

Moreover, following Singer et al. [79], our video DM is conditioned on the frame rate (fps) and we choose the times $0 \leq \tau_i \leq 1$ accordingly by sampling fps $\in \{4, 8, 12\}$ and the starting time. We render videos from the 4D scene and condition the video DM with the sampled fps. This helps generating not only sufficiently long but also temporally smooth 4D animations, as different fps correspond to long-term and short-term dynamics. Therefore, when rendering short but high fps videos they only span part of the entire length of the 4D sequence. Also see Supp. Material.

Optimizing the deformation field while supervising both with a video and image DM is crucial. The video DM generates temporal dynamics, but text-to-video DMs are not as robust as general text-to-image DMs. Including the image DM during this stage ensures stable optimization and that high visual frame quality is maintained (ablations in Sec. 4).

A crucial advantage of the disentangled two stage design is that AYG's main 4D synthesis method—the main innovation of this work—could in the future in principle also be applied to 3D objects originating from other generation systems or even to synthetic assets created by digital artists.

3.3. AYG's Score Distillation in Practice

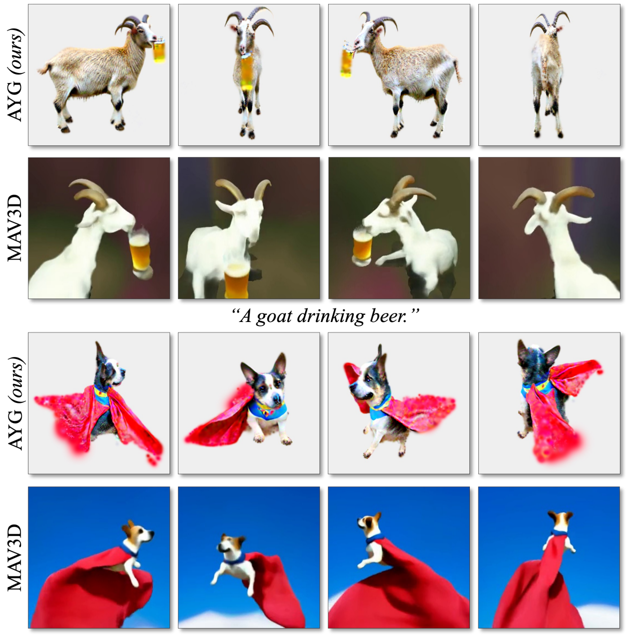
Above, we have laid out AYG's general synthesis framework. The full stage 2 score distillation gradient including CFG can be expressed as (stage 1 proceeds analogously)

$$\nabla_{\mathbf{\Phi}}\mathcal{L}^{\text{AYG}}_{\text{SDS}} = \mathbb{E}_{t,\boldsymbol{\epsilon}^{\text{vid}},\boldsymbol{\epsilon}^{\text{im}}}\Bigg[w(t)\Bigg\{\gamma\Bigg(\omega_{\text{vid}}\left[\hat{\boldsymbol{\epsilon}}^{\text{vid}}(\mathbf{Z},v,t)-\hat{\boldsymbol{\epsilon}}^{\text{vid}}(\mathbf{Z},t)\right] + \underbrace{\hat{\boldsymbol{\epsilon}}^{\text{vid}}(\mathbf{Z},v,t)-\boldsymbol{\epsilon}^{\text{vid}}}_{\boldsymbol{\delta}^{\text{vid}}_{\text{gen}}}\Bigg) + \kappa\Bigg(\omega_{\text{im}}\left[\hat{\boldsymbol{\epsilon}}^{\text{im}}(\tilde{\mathbf{Z}},v,t)-\hat{\boldsymbol{\epsilon}}^{\text{im}}(\tilde{\mathbf{Z}},t)\right] + \underbrace{\hat{\boldsymbol{\epsilon}}^{\text{im}}(\tilde{\mathbf{Z}},v,t)-\boldsymbol{\epsilon}^{\text{im}}}_{\boldsymbol{\delta}^{\text{im}}_{\text{gen}}}\Bigg)\Bigg\}\frac{\partial\{\mathbf{x}\}}{\partial\mathbf{\Phi}}\Bigg], \tag{3}$$

where $\mathbf{Z} := \{\mathbf{z}^{c_i}_{\tau_i}\}_F$, $\tilde{\mathbf{Z}} := \{\tilde{\mathbf{z}}^{\tilde{c}_j}_{\tilde{\tau}_j}\}_M$, $\omega_{\text{vid/im}}$ are the CFG scales for the video and image DMs, $\hat{\boldsymbol{\epsilon}}^{\text{vid}}(\mathbf{Z},v,t)$ and $\hat{\boldsymbol{\epsilon}}^{\text{im}}(\tilde{\mathbf{Z}},v,t)$ are the corresponding denoiser networks and $\boldsymbol{\epsilon}^{\text{vid}}$ and $\boldsymbol{\epsilon}^{\text{im}}$ are the diffusion noises (an analogous SDS gradient can be written for stage 1). Moreover, $\{\mathbf{x}\}$ denotes the set of all renderings from the 4D scene through which the SDS gradient is backpropagated, and which are diffused to produce $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$. Recently, ProlificDreamer [92] proposed a scheme where the control variates $\boldsymbol{\epsilon}^{\text{vid/im}}$ above are replaced by DMs that model the rendering distribution, are initialized from the DMs guiding the synthesis ($\hat{\boldsymbol{\epsilon}}^{\text{vid}}(\mathbf{Z},v,t)$ and $\hat{\boldsymbol{\epsilon}}^{\text{im}}(\tilde{\mathbf{Z}},v,t)$ here), and are then slowly fine-tuned on the diffused renderings ($\mathbf{Z}$ or $\tilde{\mathbf{Z}}$ here). This means that at the beginning of optimization the terms $\boldsymbol{\delta}^{\text{vid/im}}_{\text{gen}}$ in Eq. (3) would be zero. Inspired by this observation and aiming to avoid ProlificDreamer's cumbersome fine-tuning, we instead propose to simply set $\boldsymbol{\delta}^{\text{vid/im}}_{\text{gen}} = 0$ entirely and optimize with

$$\nabla_{\mathbf{\Phi}}\mathcal{L}^{\text{AYG}}_{\text{CSD}} = \mathbb{E}_{t,\boldsymbol{\epsilon}^{\text{vid}},\boldsymbol{\epsilon}^{\text{im}}}\Bigg[w(t)\Bigg\{\omega_{\text{vid}}\Big[\underbrace{\hat{\boldsymbol{\epsilon}}^{\text{vid}}(\mathbf{Z},v,t)-\hat{\boldsymbol{\epsilon}}^{\text{vid}}(\mathbf{Z},t)}_{\boldsymbol{\delta}^{\text{vid}}_{\text{cls}}}\Big] + \omega_{\text{im}}\Big[\underbrace{\hat{\boldsymbol{\epsilon}}^{\text{im}}(\tilde{\mathbf{Z}},v,t)-\hat{\boldsymbol{\epsilon}}^{\text{im}}(\tilde{\mathbf{Z}},t)}_{\boldsymbol{\delta}^{\text{im}}_{\text{cls}}}\Big]\Bigg\}\frac{\partial\{\mathbf{x}\}}{\partial\mathbf{\Phi}}\Bigg], \tag{4}$$

where we absorbed γ and κ into $\omega_{\text{vid/im}}$. Interestingly, this exactly corresponds to the concurrently proposed classifier score distillation (CSD) [102], which points out that the above two terms $\boldsymbol{\delta}^{\text{vid/im}}_{\text{cls}}$ in Eq. (4) correspond to implicit classifiers predicting v from the video or images, respectively. CSD then uses only $\boldsymbol{\delta}^{\text{vid/im}}_{\text{cls}}$ for score distillation, resulting in improved performance over SDS. We discovered that scheme independently, while aiming to inherit ProlificDreamer's strong performance. Supp. Material for details.



"A goat drinking beer."

"A dog wearing a Superhero outfit with red cape flying through the sky."

Figure 8. **AYG (*ours*) vs. MAV3D [79].** We show four 4D frames for different times and camera angles (also see Supp. Video).

3.4. Scaling Align Your Gaussians

To scale AYG and achieve state-of-the-art text-to-4D performance, we introduce several further novel techniques.

Distribution Regularization of 4D Gaussians. We developed a method to stabilize optimization and ensure realistic learnt motion. We calculate the means $\boldsymbol{\nu}_\tau$ and diagonal covariances $\boldsymbol{\Gamma}_\tau$ of the entire set of 3D Gaussians (using their means $\boldsymbol{\mu}_i$) at times τ along the 4D sequence (Fig. 4). Defining a Normal distribution $\mathcal{N}(\boldsymbol{\nu}_\tau, \boldsymbol{\Gamma}_\tau)$ with these means and covariances, we regularize with a modified version of the Jensen-Shannon divergence $\text{JSD}\left(\mathcal{N}(\boldsymbol{\nu}_0, \boldsymbol{\Gamma}_0)||\mathcal{N}(\boldsymbol{\nu}_\tau, \boldsymbol{\Gamma}_\tau)\right)$ between the 3D Gaussians at the initial and later frames τ (see Supp. Material). This ensures that the mean and the diagonal covariance of the distribution of the Gaussians stay approximately constant and encourages AYG to generate meaningful and complex dynamics instead of simple global translations and object size changes.

Extended Autoregressive Generation. By default, AYG produces relatively short 4D sequences, which is due to the guiding text-to-video model, which itself only generates short video clips (see Blattmann et al. [7]). To overcome this limitation, we developed a method to autoregressively extend the 4D sequences. We use the middle 4D frame from a first sequence as the initial frame of a second sequence, optimizing a second deformation field, optionally using a different text prompt. As the second sequence is initialized from the middle frame of the first sequence, there is an overlap interval with length 0.5 of the total length of each sequence. When optimizing for the second deformation field, we smoothly interpolate between the first and second deformation fields for the overlap region (Fig. 5). Specifically, we define $\boldsymbol{\Delta}_{\boldsymbol{\Phi}_{12}}^{\text{interpol}} = (1-\chi(\tau))\boldsymbol{\Delta}_{\boldsymbol{\Phi}_1} + \chi(\tau)\boldsymbol{\Delta}_{\boldsymbol{\Phi}_2}$ where χ is a linear function with $\chi(\tau_{0.5}) = 0$ and $\chi(\tau_{1.0}) = 1$, $\tau_{0.5}$ and $\tau_{1.0}$ represent the middle and last time frames of the first sequence, $\boldsymbol{\Delta}_{\boldsymbol{\Phi}_{12}}^{\text{interpol}}$ is the interpolated deformation field, and $\boldsymbol{\Delta}_{\boldsymbol{\Phi}_1}$(kept fixed) and $\boldsymbol{\Delta}_{\boldsymbol{\Phi}_2}$ are the deformation fields of the first and second sequence, respectively. We additionally minimize $\mathcal{L}_{\text{Interpol-Reg.}} = ||\boldsymbol{\Delta}_{\boldsymbol{\Phi}_1} - \boldsymbol{\Delta}_{\boldsymbol{\Phi}_{12}}^{\text{interpol}}||_2^2$ within the overlap region to regularize the optimization process of $\boldsymbol{\Delta}_{\boldsymbol{\Phi}_2}$. For the non-overlap regions, we just use the corresponding $\boldsymbol{\Delta}_{\boldsymbol{\Phi}}$. With this careful interpolation technique the deformation field smoothly transitions from the first sequence's into the second sequence's. Without it, we obtained abrupt, unrealistic transitions.

Table 1. **Comparison to MAV3D [79]** by user study on synthesized 4D scenes with 28 text prompts. Numbers are percentages.

Method preference	AYG *(ours)* preferred	MAV3D [79] preferred	Equal preference
Overall Quality	**53.6**	38.8	7.6
3D Appearance	**47.4**	37.2	15.4
3D Text Alignment	**45.9**	38.8	15.3
Motion Amount	**45.9**	38.8	15.3
Motion Text Alignment	**47.4**	33.7	18.9
Motion Realism	**44.4**	43.9	11.7

Motion Amplification. When a set of 4D scene renderings is given to the text-to-video model, it produces a (classifier) score distillation gradient for each frame i. We expect most motion when the gradient for each frame points into a different direction. With that in mind, we propose a motion amplification technique. We post-process the video model's individual frame scores $\boldsymbol{\delta}_{\text{cls }i}^{\text{vid}}$ $(i \in \{1, ..., F\})$ as $\boldsymbol{\delta}_{\text{cls }i}^{\text{vid}} \rightarrow \left\langle \boldsymbol{\delta}_{\text{cls }i}^{\text{vid}} \right\rangle + \omega_{\text{ma}} \left(\boldsymbol{\delta}_{\text{cls }i}^{\text{vid}} - \left\langle \boldsymbol{\delta}_{\text{cls }i}^{\text{vid}} \right\rangle \right)$, where $\left\langle \boldsymbol{\delta}_{\text{cls }i}^{\text{vid}} \right\rangle$ is the average score over the F video frames and ω_{ma} is the motion amplifier scale. This scheme is inspired by CFG and reproduces regular video model scores for $\omega_{\text{ma}}{=}1$. For larger ω_{ma}, the difference between the individual frames' scores and the average is amplified, thereby encouraging larger frame differences and more motion.

View Guidance. In AYG's 3D stage, for the text-to-image model we use a new *view guidance*. We construct an additional implicit classifier term $\omega_{\text{vg}}\left[\hat{\boldsymbol{\epsilon}}^{\text{im}}(\mathbf{z}, v^{\text{aug}}, t) - \hat{\boldsymbol{\epsilon}}^{\text{im}}(\mathbf{z}, v, t)\right]$, where v^{aug} denotes the original text prompt v augmented with directional texts such as "front view" (see Sec. 3.2) and ω_{vg} is the guidance scale. View guidance amplifies the effect of directional text prompt augmentation.

Negative Prompting. We also use negative prompt guidance during both the 3D and 4D stages. During the 4D stage, we use *"low motion, static statue, not moving, no motion"* to encourage AYG to generate more dynamic and vivid 4D scenes. Supp. Material for 3D stage and details.

4. Experiments

Text-to-4D. In Fig. 6, we show text-to-4D sequences generated by AYG (hyperparameters and details in Supp. Ma-

Align Your Gaussians (full model)	Overall Quality	3D Appearance	3D Text Alignment	Motion Amount	Motion Text Alignment	Motion Realism
v.s. w/o rigidity regularization	**45.8**/13.3	**43.3**/19.2	**38.3**/15.0	**40.8**/15.0	**42.5**/18.3	**30.8**/26.7
v.s. w/o motion amplifier	**43.3**/23.3	**37.5**/28.3	**30.8**/26.7	**45.8**/10.8	**37.5**/26.7	**33.3**/31.7
v.s. w/o initial 3D stage	**67.5**/15.0	**57.5**/21.7	**64.2**/15.0	**60.8**/21.7	**60.8**/20.8	**59.2**/24.2
v.s. w/o JSD-based regularization	**40.0**/25.0	**40.0**/27.5	**36.7**/27.5	**41.7**/24.2	**39.2**/29.2	**45.0**/24.2
v.s. w/o image DM score in 4D stage	**42.5**/22.5	**39.2**/27.5	**36.7**/25.8	**33.3**/25.9	**37.5**/30.0	27.5/**40.0**
v.s. SDS instead of CSD	**44.2**/35.8	**40.0**/27.5	**35.8**/35.0	**35.0**/27.5	**35.0**/34.2	32.5/**35.8**
v.s. 3D stage w/o MVDream	**66.7**/21.7	**48.3**/34.2	**38.3**/34.2	**41.7**/22.5	**40.0**/27.5	**40.8**/27.5
v.s. 4D stage with MVDream	**50.8**/27.5	**38.3**/34.2	**41.6**/29.2	**39.2**/35.0	**44.2**/30.0	**39.2**/31.7
v.s. video model with only fps 4	**46.7**/15.8	27.5/**36.7**	**30.0**/23.3	**36.7**/30.0	**31.7**/26.7	**32.5**/28.3
v.s. video model with only fps 12	**48.3**/29.2	**30.8**/29.2	**29.2**/28.3	**35.0**/28.3	**35.0**/30.0	**39.2**/26.7
v.s. w/o dynamic cameras	**32.5**/25.0	**32.5**/31.7	**35.0**/33.3	**35.0**/32.5	**35.8**/33.3	**32.5**/25.0
v.s. w/o negative prompting	**44.2**/28.3	**38.3**/32.5	**31.7**/29.2	29.2/**31.6**	**33.3**/30.0	**37.5**/28.3

Table 2. **Ablation study** by user study on synthesized 4D scenes with 30 text prompts. For each pair of numbers, the left number is the percentage that the full AYG model is preferred and the right number indicates preference percentage for ablated model as described in left column. The numbers do not add up to 100 and the difference is due to users voting "no preference" (details in Supp. Material).

terial). AYG can generate realistic, expressive, detailed and vivid dynamic 4D scenes (4D scenes can be rendered at varying speeds and frame rates). Importantly, our method demonstrates zero-shot generalization capabilities to creative text prompts corresponding to scenes that are unlikely to be found in the diffusion models' training images and videos. More results in Supp. Material and on project page.

To compare AYG to MAV3D [79], we performed a comprehensive user study where we took the 28 rendered videos from MAV3D's project page[2] and compared them to corresponding generations from AYG with the same text prompts (Table 1). We asked the users to rate overall quality, 3D appearance and text alignment, as well as motion amount, motion text alignment and motion realism (user study details in Supp. Material). AYG outperforms MAV3D on all metrics, achieving state-of-the-art text-to-4D performance (we also evaluated R-Precision [32, 58] on a larger prompt set used by MAV3D [78, 79], performing on par, see Supp. Mat.; however, R-Precision is a meaningless metric to evaluate *dynamic* scenes). Qualitative comparisons are shown in Fig. 8 (more in Supp. Mat.). We see that AYG produces more detailed 4D outputs. Note that MAV3D uses an extra background model, while AYG does not. Adding a similar background model would be easy but is left to future work.

Ablation Studies. Next, we performed an ablation study on AYG's different components. We used a set of 30 text prompts and generated 4D scenes for versions of AYG with missing or modified components, see Table 2. Using the same categories as before, we asked users to rate preference of our full method vs. the ablated AYG variants. Some components have different effects with respect to 3D appearance and motion, but we generally see that all components matter significantly in terms of overall quality, *i.e.*, for all ablations our full method is strongly preferred over the ablated AYG versions. This justifies AYG's design. A thorough discussion is presented in the Supp. Material, but we highlight some relevant observations. We see that our novel JSD-based regularization makes a major difference, and we also observe that the motion amplifier indeed has a strong effect for "Motion Amount". Moreover, our compositional approach is crucial. Running the 4D stage without image DM feedback produces much worse 3D and overall quality. Also the decomposition into two stages is important—carrying out 4D synthesis without initial 3D stage performs poorly.

Temporally Extended 4D Synthesis and Large Scene Composition. In Fig. 7, we show autoregressively extended text-to-4D results with changing text prompts (also see Supp. Video). AYG can realistically connect different 4D sequences and generate expressive animations with changing dynamics and behavior. We can also create sequences that loop endlessly by enforcing that the last frame of a later sequence matches the first frame of an earlier one and suppressing the deformation field there (similar to how we enforce zero deformation at τ=0 in Sec. 3.1). Finally, due to the explicit nature of the dynamic 3D Gaussians, AYG's 4D representation, multiple animated 4D objects can be easily composed into larger scenes, each shape with its own deformation field defining its dynamics. We show this in Fig. 1, where each dynamic object in the large scene is generated, except for the ground plane. These capabilities, not shown by previous work [79], are particularly promising for practical content creation applications.

5. Conclusions

We presented *Align Your Gaussians* for expressive text-to-4D synthesis. AYG builds on dynamic 3D Gaussian Splatting with deformation fields as well as score distillation with multiple composed diffusion models. Novel regularization and guidance techniques allow us to achieve state-of-the-art dynamic scene generation and we also show temporally extended 4D synthesis as well as the composition of multiple dynamic objects within a larger scene. AYG has many potential applications for creative content creation and it could also be used in the context of synthetic data generation. For example, AYG would enable synthesis of videos and 4D sequences with exact tracking labels, useful for training discriminative models. AYG currently cannot easily produce topological changes of the dynamic objects. Overcoming this limitation would be an exciting avenue for future work. Other directions include scaling AYG beyond object-centric generation and personalized 4D synthesis. The initial 3D object could be generated from a personalized diffusion model (*e.g.* DreamBooth3D [66, 71]) or with image-to-3D methods [29, 42, 44, 45, 64] and then animated with AYG.

[2] https://make-a-video3d.github.io/

References

[1] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-Shift: Latent Diffusion with Temporal Shift for Efficient Text-to-Video Generation. *arXiv preprint arXiv:2304.08477*, 2023. 2, 3, 15

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016. 17

[3] 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. 16

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 4, 24

[5] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-Image Diffusion Models with Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 3, 15

[6] Miguel Ángel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander T Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, Afshin Dehghan, and Joshua M. Susskind. GAUDI: A Neural Architect for Immersive 3D Scene Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 15

[7] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 4, 7, 15, 19, 24, 25

[8] Hongrui Cai, Wanquan Feng, Xuetao Feng, Yan Wang, and Juyong Zhang. Neural Surface Reconstruction of Dynamic Scenes with Monocular RGB-D Camera. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3, 16

[9] Ang Cao and Justin Johnson. HexPlane: A Fast Representation for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 15

[10] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 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)*, 2023. 2, 3, 15

[11] Yiwen Chen, Chi Zhang, Xiaofeng Yang, Zhongang Cai, Gang Yu, Lei Yang, and Guosheng Lin. IT3D: Improved Text-to-3D Generation with Explicit View Synthesis. *arXiv preprint arXiv:2308.11473*, 2023. 2, 3, 15

[12] Zilong Chen, Feng Wang, and Huaping Liu. Text-to-3D using Gaussian Splatting. *arXiv preprint arXiv:2309.16585*, 2023. 3, 16

[13] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. LucidDreamer: Domain-free Generation of 3D Gaussian Splatting Scenes. *arXiv preprint arXiv:2311.13384*, 2023. 16

[14] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, Matthew Yu, Abhishek Kadian, Filip Radenovic, Dhruv Mahajan, Kunpeng Li, Yue Zhao, Vladan Petrovic, Mitesh Kumar Singh, Simran Motwani, Yi Wen, Yiwen Song, Roshan Sumbaly, Vignesh Ramanathan, Zijian He, Peter Vajda, and Devi Parikh. Emu: Enhancing Image Generation Models Using Photogenic Needles in a Haystack. *arXiv preprint arXiv:2309.15807*, 2023. 3, 15

[15] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A Universe of 10M+ 3D Objects. *arXiv preprint arXiv:2307.05663*, 2023. 4

[16] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A Universe of Annotated 3D Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[17] C. Deng, C. Jiang, C. R. Qi, X. Yan, Y. Zhou, L. Guibas, and D. Anguelov. NeRDi: Single-View NeRF Synthesis with Language-Guided Diffusion as General Image Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 15

[18] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, 2021. 3, 15

[19] Yilun Du, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Grathwohl. Reduce, Reuse, Recycle: Compositional Generation with Energy-Based Diffusion Models and MCMC. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. 3, 16, 20

[20] Lincong Feng, Muyu Wang, Maoyu Wang, Kuo Xu, and Xiaoli Liu. MetaDreamer: Efficient Text-to-3D Creation With Disentangling Geometry and Texture. *arXiv preprint arXiv:2311.10123*, 2023. 15

[21] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun, Li Chen, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-ViLG 2.0: Improving Text-to-Image Diffusion Model With Knowledge-Enhanced Mixture-of-Denoising-Experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 15

[22] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve Your Own Correlation: A Noise Prior for Video Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 15

[23] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu Video: Factorizing Text-to-Video Generation by Explicit Image Conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 15

[24] Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Josh Susskind, and Navdeep Jaitly. Matryoshka Diffusion Models. *arXiv preprint arXiv:2310.15111*, 2023. 15

[25] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3, 15

[26] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3, 23

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3, 15, 21, 22

[28] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen Video: High Definition Video Generation with Diffusion Models. *arXiv preprint arXiv:2210.02303*, 2022. 2, 3, 15

[29] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large Reconstruction Model for Single Image to 3D. *arXiv preprint arXiv:2311.04400*, 2023. 8

[30] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. Simple Diffusion: End-to-End Diffusion for High Resolution Images. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. 15

[31] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. DreamTime: An Improved Optimization Strategy for Text-to-3D Content Creation. *arXiv preprint arXiv:2306.12422*, 2023. 2, 3, 15

[32] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel, and B. Poole. Zero-Shot Text-Guided Object Generation with Dream Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8, 28

[33] 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. 16

[34] Nikolai Kalischek, Torben Peters, Jan D. Wegner, and Konrad Schindler. Tetrahedral Diffusion Models for 3D Shape Generation. *arXiv preprint arXiv:2211.13220*, 2022. 15

[35] Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. Noise-Free Score Distillation. *arXiv preprint arXiv:2310.17590*, 2023. 15

[36] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 3, 4, 16, 17

[37] Isaac Kerlow. *The Art of 3D Computer Animation and Effects*. Wiley Publishing, 4th edition, 2009. 16

[38] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. *arXiv preprint arXiv:2303.13439*, 2023. 3, 15, 25

[39] Seung Wook Kim, Bradley Brown, Kangxue Yin, Karsten Kreis, Katja Schwarz, Daiqing Li, Robin Rombach, Antonio Torralba, and Sanja Fidler. NeuralField-LDM: Scene Generation with Hierarchical Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 15

[40] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. LucidDreamer: Towards High-Fidelity Text-to-3D Generation via Interval Score Matching. *arXiv preprint arXiv:2311.11284*, 2023. 15, 16

[41] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3D: High-Resolution Text-to-3D Content Creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 15

[42] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast Single Image to 3D Objects with Consistent Multi-View Generation and 3D Diffusion. *arXiv preprint arXiv:2311.07885*, 2023. 3, 8, 15

[43] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional Visual Generation with Composable Diffusion Models. In *Computer Vision – ECCV 2022*, 2022. 3, 16, 20

[44] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3, 8, 15

[45] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. *arXiv preprint arXiv:2309.03453*, 2023. 3, 8, 15

[46] Zhen Liu, Yao Feng, Michael J. Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. MeshDiffusion: Score-based Generative 3D Mesh Modeling. In *International Conference on Learning Representations (ICLR)*, 2023. 15

[47] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3D: Single Image to 3D using Cross-Domain Diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 15

[48] Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa, Nicholas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, and James Lucas. ATT3D: Amortized Text-to-3D Object Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3, 15

[49] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, 1981. 24

[50] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 3, 4, 16, 17

[51] Shitong Luo and Wei Hu. Diffusion Probabilistic Models for 3D Point Cloud Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 15

[52] G. Metzer, E. Richardson, O. Patashnik, R. Giryes, and D. Cohen-Or. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 15

[53] Marko Mihajlovic, Sergey Prokudin, Marc Pollefeys, and Siyu Tang. ResFields: Residual Neural Fields for Spatiotemporal Signals. *arXiv preprint arXiv:2309.03160*, 2023. 16

[54] 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. 3, 15

[55] Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 17

[56] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. *arXiv preprint arXiv:2212.08751*, 2022. 3, 15

[57] Alexander Quinn Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. 3, 15

[58] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for Compositional Text-to-Image Synthesis. In *NeurIPS Datasets and Benchmarks*, 2021. 8, 28

[59] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 4, 16

[60] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields. *ACM Trans. Graph.*, 40(6), 2021. 3, 16

[61] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3, 15

[62] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D Diffusion. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 5, 15, 16, 21, 22

[63] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 4, 16

[64] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors. *arXiv preprint arXiv:2306.17843*, 2023. 3, 8, 15

[65] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. 28

[66] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Ben Mildenhall, Nataniel Ruiz, Shiran Zada, Kfir Aberman, Michael Rubenstein, Jonathan Barron, Yuanzhen Li, and Varun Jampani. DreamBooth3D: Subject-Driven Text-to-3D Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 8

[67] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*, 2022. 3, 15

[68] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and Pace: Controllable Pedestrian Animation via Guided Trajectory Diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 16

[69] Geoffrey Roeder, Yuhuai Wu, and David Duvenaud. Sticking the Landing: Simple, Lower-Variance Gradient Estimators for Variational Inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 22

[70] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4, 15, 19, 20, 24, 29

[71] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[72] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed

Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 15

[73] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 24

[74] Katja Schwarz, Seung Wook Kim, Jun Gao, Sanja Fidler, Andreas Geiger, and Karsten Kreis. WildFusion: Learning 3D-Aware Latent Diffusion Models in View Space. *arXiv preprint arXiv:2311.13570*, 2023. 15

[75] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a Single Image to Consistent Multi-view Diffusion Base Model. *arXiv preprint arXiv:2310.15110*, 2023. 3, 15

[76] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. MVDream: Multi-view Diffusion for 3D Generation. *arXiv preprint arXiv:2308.16512*, 2023. 2, 3, 4, 15, 19, 20, 24, 25, 29

[77] J. Shue, E. Chan, R. Po, Z. Ankner, J. Wu, and G. Wetzstein. 3D Neural Field Generation Using Triplane Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 15

[78] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 8, 15, 16, 24, 27, 28

[79] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, and Yaniv Taigman. Text-to-4D Dynamic Scene Generation. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. 2, 3, 4, 6, 7, 8, 15, 17, 20, 24, 26, 27, 30, 31, 36

[80] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015. 3, 15

[81] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*, 2021. 3, 15, 21, 22

[82] Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. DreamCraft3D: Hierarchical 3D Generation with Bootstrapped Diffusion Prior. *arXiv preprint arXiv:2310.16818*, 2023. 15

[83] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. *arXiv preprint arXiv:2309.16653*, 2023. 3, 16, 25

[84] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-Rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Dynamic Scene From Monocular Video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3, 16

[85] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. TextMesh: Generation of Realistic 3D Meshes From Text Prompts. In *International conference on 3D vision (3DV)*, 2024. 2, 3, 15

[86] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based Generative Modeling in Latent Space. In *Neural Information Processing Systems (NeurIPS)*, 2021. 4, 15, 22, 24

[87] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in neural information processing systems*, 30, 2017. 17

[88] 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 (CVPR)*, 2023. 2, 3, 15

[89] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. RODIN: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 15

[90] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. VideoFactory: Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. *arXiv preprint arXiv:2305.10874*, 2023. 2, 3, 4, 15, 24

[91] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. LAVIE: High-Quality Video Generation with Cascaded Latent Diffusion Models. *arXiv preprint arXiv:2309.15103*, 2023. 2, 3, 15

[92] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3, 6, 15, 23

[93] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. *arXiv preprint arXiv:2310.08528*, 2023. 3, 16

[94] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu

Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3, 15, 25

[95] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. PhysGaussian: Physics-Integrated 3D Gaussians for Generative Dynamics. *arXiv preprint arXiv:2311.12198*, 2023. 16

[96] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. NeuralLift-360: Lifting An In-the-wild 2D Photo to A 3D Object with 360° Views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 15

[97] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, and Kai Zhang. DMV3D: Denoising Multi-View Diffusion using 3D Large Reconstruction Model. *arXiv preprint arXiv:2311.09217*, 2023. 15

[98] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. RAPHAEL: Text-to-Image Generation via Large Mixture of Diffusion Paths. *arXiv preprint arXiv:2305.18295*, 2023. 3, 15

[99] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. BANMo: Building Animatable 3D Neural Models from Many Casual Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 16

[100] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction. *arXiv preprint arXiv:2309.13101*, 2023. 16

[101] Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. GaussianDreamer: Fast Generation from Text to 3D Gaussian Splatting with Point Cloud Priors. *arXiv preprint arxiv:2310.08529*, 2023. 3, 16

[102] Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3D with Classifier Score Distillation. *arXiv preprint arXiv:2310.19415*, 2023. 2, 4, 6, 16, 23, 24

[103] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-Guided Human Motion Diffusion Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 16

[104] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. LION: Latent Point Diffusion Models for 3D Shape Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 15

[105] Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating One Image to 4D Dynamic Scene. *arXiv preprint arXiv:2311.14603*, 2023. 16

[106] Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Karsten Kreis, Otmar Hilliges, and Shalini De Mello. A Unified Approach for Text- and Image-guided 4D Scene Generation. *arXiv preprint arXiv:2311.16854*, 2023. 16

[107] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. MagicVideo: Efficient Video Generation With Latent Diffusion Models. *arXiv preprint arXiv:2211.11018*, 2023. 2, 3, 15

[108] Linqi Zhou, Yilun Du, and Jiajun Wu. 3D Shape Generation and Completion Through Point-Voxel Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 15

[109] Junzhe Zhu and Peiye Zhuang. HiFA: High-fidelity Text-to-3D with Advanced Diffusion Guidance. *arXiv preprint arXiv:2305.18766*, 2023. 2, 3, 15

[110] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3D Gaussian Avatars. *arXiv preprint arxiv:2311.08581*, 2023. 3, 16

[111] M. Zwicker, H. Pfister, J. van Baar, and M. Gross. EWA volume splatting. In *Proceedings Visualization, 2001. VIS '01.*, 2001. 3