HarmonyDreamer: Rectifying and Harmonizing Dynamic Objects in Backgrounds for 4D Video Generation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent advancements in score distillation have markedly enhanced the generation of both static and dynamic 3D tasks by effectively harnessing the power of 2D pretrained diffusion models for artistic applications. However, traditional distillation methods primarily concentrate on object-centric representations without adequately addressing the integration of objects with their backgrounds, which includes matching rendering styles and motions. This paper presents a new framework that incorporates a background-enhanced distillation score along with an iterative editing-reconstruction process tailored for 4D video generation from basic dynamic 4D Gaussian splatting. Our experiments demonstrate that our method not only excels in producing high-fidelity 4D videos that harmoniously blend objects with varied backgrounds, but also facilitates the creation of multi-view videos. Additionally, our approach outperforms existing baselines in both spatio-temporal and semantic consistency, demonstrating its effectiveness in 4D video generation. More video results can be viewed in our project page: `https://harmonydreamer.github.io/web/`.

1 Introduction

Recent advancements in generative models, particularly text-to-image and text-to-video diffusion models, have paved the way for a new era of content creation. Studies show that these models have achieved remarkable success in generating various digital formats, encompassing 2D images [34, 31, 30], videos [3, 15, 5], and 3D scenes [28, 43, 40]. Beyond producing contents from above-mentioned modalities, considerable effort has been directed towards developing more sophisticated content such as videos and 4D scenes. The majority of research focuses on text-to-4D [36, 2, 46, 24], video-to-4D [8, 33, 49], and image-to-4D approaches [33, 51, 47]. However, most of these methods are far from practical usage as they can only do object-centric generation without integrating these objects under arbitrary background with matching rendering styles, shadows, and motion.

In this paper, we introduce the inversion score for both image and video harmonization. We also propose a workflow as illustrated in Fig. 1 that processes rough 3D or 4D inputs for multi-view video generation. Our pipeline begins by stacking rough Gaussian splatting 3D inputs onto a layered image background. We then distill the background semantics to the front 3D input, enabling both rectification and harmonization using our proposed image layered inversion score. Following this, the edited 3D result can be attached to a motion network to produce a rough 4D output, which can be further rectified using our proposed video inversion score.

In the text-to-3D and editing tasks, score-based distillation [28] methods have been extensively studied to distill object-centric 3D objects that align with monocular views or text prompts. Another

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

approach, starting with Instruct-Nerf2Nerf [13], proposes a 3D scene editing method that continually trains a 3D representation while editing partial multi-view rendering images using off-the-shelf image editing method InstructPix2Pix [4]. This variation of the score distillation method can edit more general 3D scenes with high resolution but is time-consuming and limited to text-guided editing.

To harmonize visual artifacts and background semantics, a distillation method must be flexible, not constrained to object-centric information, and robust against unsatisfactory inputs with layered artifacts and camera intrinsic inconsistencies. Our method demonstrates that the text-to-image off-the-shelf model can perform harmonization using layered inversion. Additionally, we show that the same scheme works for video inversion, enabling 4D harmonization, including rectification and interaction dynamics discovery. Our contributions can be listed as follows:

- We propose a versatile 3D editing method for harmonization and rectification that is not constrained to object-centric scenes.
- We explore the semantics derived from layered backgrounds, in addition to those from text prompts.
- We develop an inversion score that is applicable to both images and videos, extending 3D harmonization results to 4D with arbitrary input motions. This allows for the decoupling of editing appearance and motion style in multi-view video creation.

2 Related Work

Text-to-3D Generation The limited availability of extensive text-annotated 3D training datasets hampers progress in 3D content generation. To overcome such limitations, DreamField [17] combined neural rendering [27] and multi-modal representation [29] to directly generate 3D assets without 3D supervision. However, most of the recent advancement utilized the pretrained Text-to-Image models such as Stable Diffusion [34] to serve as a strong generative prior for text-to-3D generation. Dreamfusion [28] pioneered such pipeline by translating these image-based priors into coherent 3D models with an iterative distillation approach, named as Score Distillation Sampling (SDS). It leverages and back-propagates random view samples from a 2D diffusion model to optimize 3D representations such as NeRF [27], through a mode seeking manner. Sparked by the novel SDS loss, research in this area has concentrated on several key objectives. Several of them aim to introducing better update rule for score distillation to overcome the artifacts in the generated 3D assets such as the low-diversity, over-saturation, and the *Janus* problem [43, 42, 41, 23, 52], while the other choose to optimize over more efficient 3D representation such as Gaussian Splatting [22] to accelerate the content generation process [40, 48, 7, 11].

4D Generation 4D content Generation targets to generate dynamic and spatial-temporal consistent 3D scenes viewed at arbitrary angles. While previous methods focus on generating dynamic object of specific classes such as animals [18] and humans [32, 26, 50], Make-A-Video3D (MAV3D) [36] explores this area by extending the SDS-based approach with 4D Dynamic NeRF and HexPlane representation [6] to generate arbitrary objects and flexibly model large scene motions. Consistent4D [20] proposes the interpolation-driven consistency loss with per-frame SDS to generate dynamic objects under uncalibrated monocular videos. 4Dfy [2] blends the SDS supervision from both Text-to-image and Text-to-video diffusion models to alleviates the tradeoff between appearance, 3D structure, and motion in 4D assets generation. 4DGen [49] combined DDIM-sampled pseudo labels with SDS to ensure spatial-temporal consistent and high-resolution generation. TC4D [1] proposed a trajectory-conditioned method to supervise the global and local scene animation individually via video score distillation sampling. Most of the existing works focus on object-centric generation without background information. Our work, however, aims to fill the gap by incorporating a novel and enhanced background information distillation score.

3 Method

3.1 Preliminaries

Sampling Method Diffusion-based generative models, applicable to both images and videos, utilize a score matching function for training. These models can be sampled using expensive stochastic methods, such as the Denoising Diffusion Probabilistic Models (DDPM) [16]. An advancement in this field, the Denoising Diffusion Implicit Models (DDIM) [37], accelerates the sampling process by

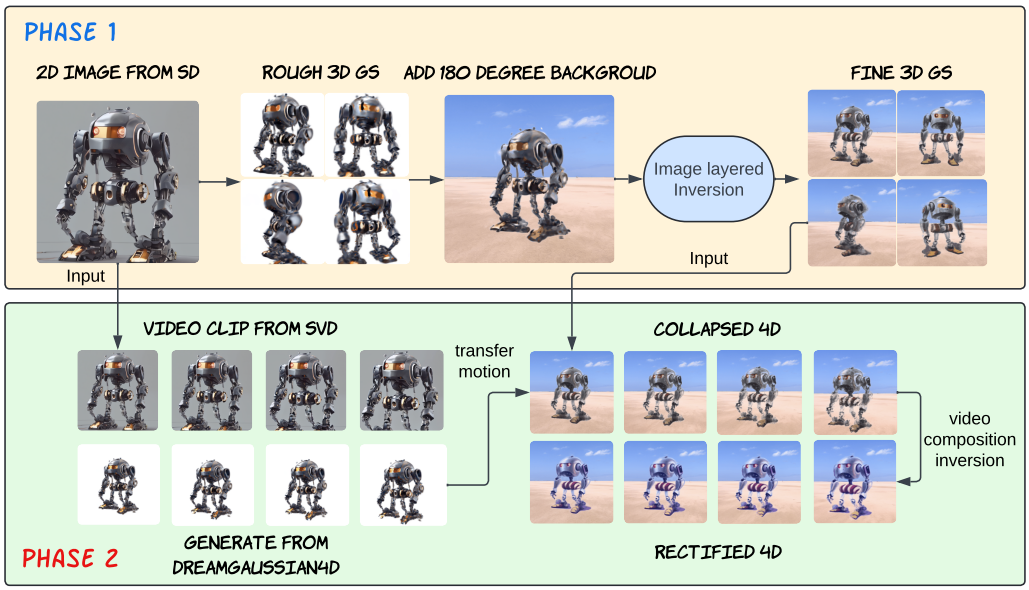

Figure 1: Our Method contains two phases. Phase 1 takes a 2D generated model from Stable diffusion, generate the low-quality 3D Gaussian splatting model and refine it with by adding background and our image layered inversion. Phase 2 takes the generated 2D image to get the 3D reference video from Stable Video diffusion. Then we use DreamGaussian4D to learn the motion from the video and transfer to the refined 3D object and further improve the harmonization with our video composition inversion to get the final output.

conceptualizing it as an Ordinary Differential Equation (ODE) probability flow. This method has gained popularity as the predominant sampling technique for pre-trained image generation models. Further developments include the Enhanced Diffusion Model (EDM), which explains both Stochastic-Multiscale-Langevin Dynamics (SMLD) [38] and DDIM within a single ODE sampling framework, allowing for flexible selection of sampling time steps. Owing to its continuous nature, the EDM schedule proves to be particularly well-suited for Stable Video Diffusion and is thus commonly employed as the default method.

During the inference phase of text-to-image diffusion model, the deterministic sampling of DDIM [37] follows an ODE-like deterministic trajectory:

$$\frac{z_{t-1}}{\sqrt{\alpha_{t-1}}} = \frac{z_t}{\sqrt{\alpha_t}} + \left(\frac{\sqrt{1-\alpha_{t-1}}}{\sqrt{\alpha_{t-1}}} - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}} \right) \epsilon_\theta(z_t, t) \tag{1}$$

And during the inference phase of image-to-video diffusion model, the deterministic sampling of EDM is as follows:

$$z_{t-1} = z_t + (\sigma(t-1) - \sigma(t)) \frac{D(z; \sigma(t))}{\sigma(t)}, \tag{2}$$

where z_t represents the latent sampled at time step t, α_t and $\sigma(t)$ are the time-dependent diffusion coefficients, and ϵ_θ represents the denoiser module parameterized by a neural network. We follow the same convention in SVD default sampling, setting $\sigma(t) = t$ and $s(t) = 1$.

Gaussian Splatting and Dynamic Gaussian splatting 3D Gaussian Splatting [22] aims to reconstruct a 3D scene by initializing a set of 3D Gaussian primitives and minimizing the photometric discrepancy between a set of input images $\{I\}_m$ paired with known camera views $\{V\}_m$ and their corresponding rendered images $\{I_r\}_m$. In practice, we denote one Gaussian splatting representation as G, with G^t being the Gaussian splatting at time frame t. Here, G_c^t represents colors, G_s^t denotes the scales, G_r^t indicates the rotations, and $G_{\mathbf{x}}^t$ signifies the Gaussian splats' position centers. We denote $G(v)$ as the rendered image for a given camera view v and $G_{\mathbf{a}}(v)$ as the corresponding alpha mask. For simplicity, in the following text, we will abbreviate Gaussian Splatting as GS.

3.2 Inversion Score

3.2.1 ODE Sampling Inversion

z_0 represents the encoded form of the provided real image or video. Rather than the typical z_T to z_0 sampling, the inversion process operates inversely, transitioning from z_0 to z_T. z_T can be used to extract rough semantic information for further editing purposes. To leverage the inverse characteristics of the ODE path, we can apply a similar inversion scheme to the DDIM[37] sampler within the text-to-image model:

$$\frac{z_{t+1}}{\sqrt{\alpha_{t+1}}} = \frac{z_t}{\sqrt{\alpha_t}} + \left(\frac{\sqrt{1-\alpha_{t+1}}}{\sqrt{\alpha_{t+1}}} - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}} \right) \epsilon_\theta(z_t, t, C), \tag{3}$$

and EDM[21] sampler with image-to-video model:

$$z_{t+1} = z_t + (\sigma(t+1) - \sigma(t)) \frac{D(z; \sigma(t), C)}{\sigma(t)}. \tag{4}$$

3.2.2 Explicit Modified Inversion

To streamline the notation in our discussion, we use $Inv(I, C_{text}, t(r))$ to denote the inversion process applied to an image I, with the textual condition C_{text}, and reversing to the fraction $t = r \cdot T$, r being the ratio of inversion steps completed over T being the scheduled total sampling steps. In a similar vein, $Inv(V, C_{image}, t(r)))$ refers to the inversion process for a video sample V, this time with an image-based condition C_{image}, and using EDM inversion.

Inversion process offers additional advantages [14]. [9] applies the clean data estimation $z_{0|t} = \frac{1}{\sqrt{\alpha_t}} \left(z_t - \sqrt{1-\alpha_t} \epsilon_\theta(z_t, t) \right)$ from Tweedie's formula [10] to obtain a point estimate of the noisy data. In our case, we use $Inv(I, C_{text}, t)$ to denote the inversed latents with t inversion steps, and further $Inv(I, C_{text}, t)_{0|t}$ to denote the clean data estimation from this inversed latents by $Inv(I, C_{text}, t)_{0|t} = \frac{1}{\sqrt{\alpha_t}} \left(Inv(I, C_{text}, t) - \sqrt{1-\alpha_t} \epsilon_\theta(Inv(I, C_{text}, t), t) \right)$. In Fig. 2, we present the visualizations of clean data estimation from the inversed latents, demonstrating the capture of basic geometric information from the input sample. This serves as guidance for both the reversion path and the subsequent 3D reconstruction.

Our reversion steps modify the direct sampling method in Eq. 1 and Eq. 2 as follows

$$\hat{z}_{0|t} = \mathcal{K}(t) \cdot z_{0|t} + (1 - \mathcal{K}(t)) \cdot Inv(I, C_{text}, t)_{0|t} \tag{5}$$

$$\hat{\epsilon}_\theta(z_t, t, C, \emptyset) = w \cdot \epsilon_\theta(z_t, t, C) + (1 - w) \cdot \epsilon_\theta(z_t, t, \emptyset) \tag{6}$$

$$\hat{z}_t = \sqrt{\alpha_t} \cdot \hat{z}_{0|t} + \sqrt{1-\alpha_t} \hat{\epsilon}_\theta(z_t, t, \emptyset) \tag{7}$$

$$\frac{z_{t+1}}{\sqrt{\alpha_{t+1}}} = \frac{\hat{z}_t}{\sqrt{\alpha_t}} + \left(\frac{\sqrt{1-\alpha_{t+1}}}{\sqrt{\alpha_{t+1}}} - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}} \right) \hat{\epsilon}_\theta(z_t, t, C), \tag{8}$$

Note that, as DDIM is a special case of EDM, we use DDIM notation to illustrate the modified reversion sampling. ω is the default Classifier-free Guidance [16] parameter in sampling process. $\mathcal{K}(t)$ is a decreasing weight acts as a factor that blends the inversion and reversion paths using the recorded clean data estimated during inversion. We also present visualizations of the clean data estimation along the reversion path to demonstrate its effect. For simplicity, we denote the process of iterating the modified reversion sampling using Eqs. 5, 6, and 8 from timestep t to 0 as $\text{Rev}(I, C, \omega, \mathcal{K})$. Illustrated in Fig. 2, using the first image on the left of the first row as input, the visuals of $\text{Rev}(I, C, \omega, \mathcal{K})$ would be the first image on the left in the second row. In practice, we use 15 inversion steps out of 50 total DDIM steps for the text-to-image model, and 10 inversion steps out of 25 total steps for the image-to-video model.

3.3 Image Layered Inversion

For image inversion, we use a cropped object, layer it over a background image, and apply the inversion method as described. In our practice, this object could be the rendered Gaussian Splatting and its mask. As shown in Fig. 1, the rendered robot model layered on the background demonstrates

rectification. In Fig. 5, stacking the same rough 3D input onto the background with the same text prompt results in semantically different outcomes after iterations, with the initial rough model rectified and harmonized. Fig. 7 illustrates the gradual evolution of the editing results, highlighting our method's stability and advantages for downstream applications.

3.4 Video Composition Inversion

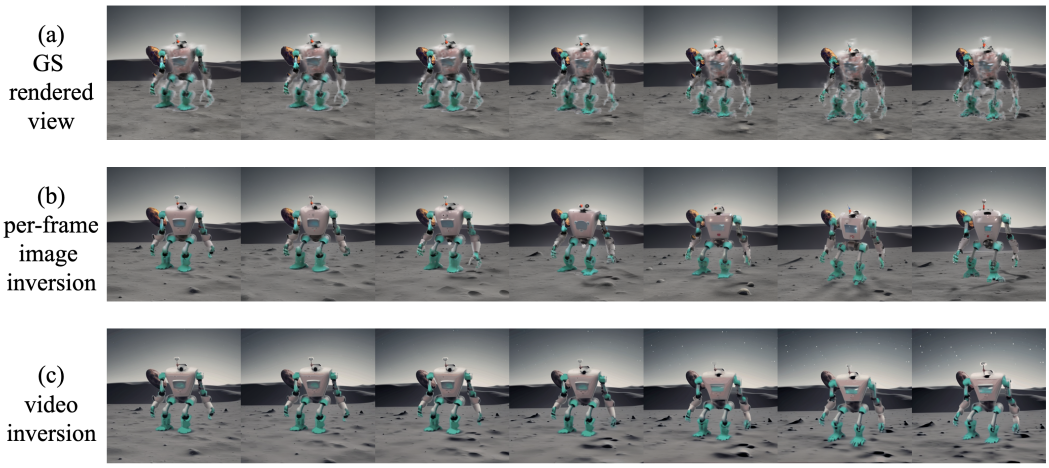

Figure 3: (a) Rendered view of temporal Gaussian Splatting layered with user-specified background, showing 7 frames extracted in sequence from a total of 14 frames. (b) Per-frame image harmonization result using layered image inversion as described in Section 3.3. (c) Temporal harmonization result using video composition inversion as described in Section 3.4.

Our inversion method can also be applied to the sampler in Stable Video Diffusion[3]. For clarity, we slightly extend our notation, using $\text{Rev}(I, C_{image})$ to represent the video inversion and reversion process. However, directly applying video inversion to rendered frames results in deteriorated quality. To address this problem, we propose **video composition inversion**. First, we apply image inversion to the rendered frames as shown in Fig. 3(a), using image layered inversion to obtain Fig. 3(b). Then, we apply video inversion to the intermediate output to achieve Fig. 3(c). Our approach effectively balances the temporal consistency and frame-wise quality.

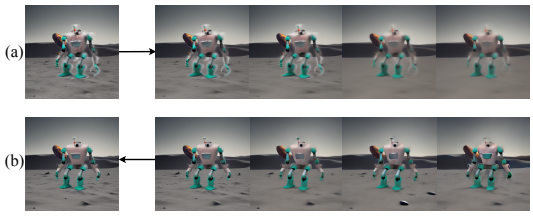

Figure 2: (a) Visuals of the image inversion path (b) Visuals of our modified reversion path.

3.5 Scheduled Optimization

Our method initiates with a preliminary Gaussian splatting G and a predefined set of views, delineated by specific ranges of horizon and azimuth. As discussed in Section 3.3, our inversion score incorporates semantics from both text prompts and background information. We stack view v on the mapped background as $I_b(v)$. For multi-view consistency, the mapping from view v to background image I_b should be continuous, though not necessarily with accurate 3D geometry, as our harmonization steps will correct any layered artifacts. We propose a simple solution by cropping a small image from a larger one with changes in horizon and azimuth.

3.5.1 Scheduled Gaussian Splatting Optimization

We denote raw input without any editing denoted as $G(i)$ for $i = 0$, where i represents the i-th Gaussian splatting after i iterations of editing. A single editing step is described as the following:

We first render n views $G(i, v)$, $v \in V$, randomly from the predefined visible range, and then layer the corresponding mapped background to each view. For example, given view v, the layered rendering

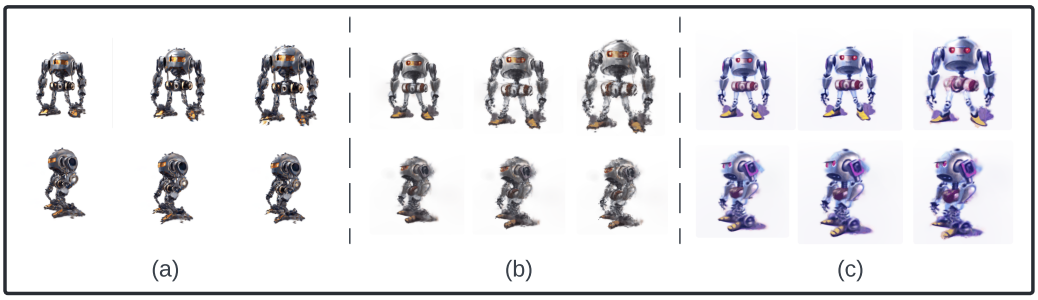

Figure 4: Quality comparison for three frames and two different views. (a) Dreamgaussian4D output. (b) Motion transfer to refined 3D. (c) Rectified 4D with our method.

image would be:

$$I_{\text{layer}}(i,v) = G_{\mathbf{a}}(v) \cdot G(i,v) + (1 - G_{\mathbf{a}}(v)) \cdot I_b(v).$$

Next, we apply inversion to m randomly selected views out of n. For simplicity, we denote the first m views as the selected ones for inversion $v = 0, 1, \ldots, m$. Therefore, the optimization objective for $G^t_{\mathbf{a}}(i,v)$ can be formulated as:

$$\begin{aligned} L = \min_{G(i)} \sum_{v=1}^{m} \| (G_{\mathbf{a}}(i,v) \cdot G(i,v) + (1 - G_{\mathbf{a}}(v)) \cdot I_b(v)) - \text{Rev}(I_{\text{layer}}(i,v), C_{\text{Text}}, t(r(v))) \|^2 \\ + \sum_{v=m+1}^{n} \| (G_{\mathbf{a}}(i,v) \cdot G(i,v) + (1 - G_{\mathbf{a}}(v)) \cdot I_b(v))) - I_{\text{layer}}(i,v) \|^2. \end{aligned} \tag{9}$$

Note that as inversion proceeds, the editing effect applied to the object becomes more divergent from the input. Due to the orbit camera intrinsics, when r is larger, the projected object appears smaller. To adapt and balance the editing effect across different views, we use the empirical formula $r(v) = 0.7 - \text{radius}(v)/10$, where $\text{radius}(v)$ is the radius of view v.

However, continuous optimization of a fixed representation can make the Gaussian splatting hard to converge and prone to degradation, preventing recovery in the editing process. Therefore, we propose a rescheduled optimization pipeline for Gaussian splatting reconstruction, which shows robustness in handling multi-view images with subtle inconsistencies, making it suitable for our inversion score. The optimization update step is described as follows:

$$G(i+1)^{(s)} = \begin{cases} G(i)^{(s)} + \mathcal{J}(s) G(i) & \text{if } s = 0 \\ G(i+1)^{(s)} + \mathcal{J}(s) G(i+1) & \text{if } s > 0 \end{cases} \tag{10}$$

$$G(i+1)^{(s+1)} = G(i+1)^{(s)} - \eta \nabla_{G(i+1)} L \tag{11}$$

We use $\mathcal{J}(s)$ as a gradually decreasing factor to control the noise added to the optimization. Note that when $s = 0$, the noise is added to the Gaussian splatting $G(i)$ from the previous editing iteration i.

Here, we use $G(i+1)$ to represent all learnable attributes in the $i+1$-th Gaussian splatting. In practice, these attributes can include colors G^t_c only, or all learnable attributes: colors $G^t_c(i+1)$, scales $G^t_s(i+1)$, rotations $G^t_r(i+1)$, and point centers $G^t_{\mathbf{x}}(i+1)$.

In the experiment, we set $n = 24$, $m = 4$, and the optimization steps for each editing iteration to 400. After this optimization, we denote the updated Gaussian splatting as $G(i+1)$. A new image set is then rendered from $G(i+1)$, and the process repeats.

3.5.2 Scheduled Dynamic Gaussian Splatting Optimization

Note that we are not utilizing any additional temporal adjustment structures such as *Gaussian-Flow*[12] or *Gaussian4D*[44]. Instead, we adopt per-frame independent Gaussian splatting as our spatio-temporal representation, denoted as $\{G^t\}$. We use $G^t(v)$ to represent the rendered image

obtained from a given camera view v at frame t. Temporal consistency is derived from video composition inversion (see Section 3.4). The term Rev^t represents the rectified frame at time t, which is subsequently passed to the Gaussian Splatting rendered loss. The objective for the stacked temporal representation is modified from Equation 9. To capture the dynamics of the layered background, we employ SVD[3] to obtain $I^t_{\text{layer}}(i, v)$.

$$L_{\text{temporal}} = \min_{\{G^t(i)\}} \sum_{v=1}^{m} \sum_{t=1}^{f} \Big\| (G^t_{\text{a}}(i,v) \cdot G^t(i,v) + (1 - G^t_{\text{a}}(v)) \cdot I_b(v)) \tag{12}$$

$$- \text{Rev}^t(\{\text{Rev}(I^t_{\text{layer}}(i,v), C_{\text{Text}}, r(v) \cdot T)\}, \text{Rev}(I^0_{\text{layer}}(i,v), C_{\text{Text}}, r(v) \cdot T)) \Big\|^2$$

$$+ \sum_{v=m+1}^{n} \Big\| (G^t_{\text{a}}(i,v) \cdot G^t(i,v) + (1 - G^t_{\text{a}}(v)) \cdot I^t_b(v)) - I^t_{\text{layer}}(i,v) \Big\|^2 . \tag{13}$$

4 Experiments

4.1 3D harmonization

First, following the initial stage settings of DreamGaussian4D [33], we use a reference image to generate a rough 3D Gaussian Splatting using reconstruction loss and distillation gradients from a view-conditioned diffusion model. Artifacts, such as rough Gaussian blobs, are visible in these initial models, as shown in Fig. 1 and the first column of Fig. 5. To align with the DreamGaussian output, we use orbit cameras (see Section A.1) and define the visible range.

Then, as explained in Section 3.5.1, we apply the scheduled editing process. The iterative editing results are shown in Fig. 7. Additionally, diverse editing results with the same prompt but different backgrounds are displayed in Fig. 5.

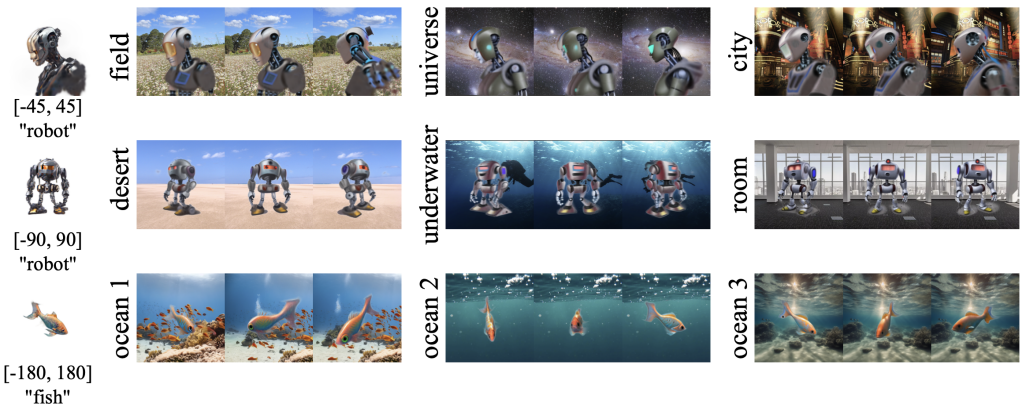

Figure 5: Editing results of input rough GS with three predefined views, harmonized to diverse appearances in different backgrounds.

4.2 4D Harmonization

Using the same reference image, the reference motion can be prompted from SVD [3]. The video reconstruction loss and view-conditioned distillation score produce a rough 4D model, as shown in Fig. 4(a). This input model is represented as a first frame GS with a deformation network [44], predicting Gaussian blob attribute increments relative to the first frame. This backbone preserves temporal consistency but is challenging to edit and converge. The deformation network accepts a modified first frame GS, including variations in Gaussian blob attributes and the number of Gaussian blobs. However, transferring motion to the edited first frame causes issues, as temporal consistency degrades over time, as shown in Fig. 4(b). In our workflow, we use the edited first frame GS with the deformation network (Fig. 4(a)), and the output (Fig. 4(b)) serves as initialization. Temporal

consistency rectification is performed using our method proposed in Section 3.5.2. Results in Fig. 7 show that even after 280 iterations, our method effectively attaches different motions to the edited character while maintaining appearance consistency. Fig. 6 demonstrates that our method can attach different first-frame GS to the same motion, serving as a flexible framework for appearance and motion editing and disentanglement. Emergent temporal semantics are observed after harmonization. For example, in Fig. 6(b), the flowers wave together, and in Fig. 10(a) in Section A, we see light reflection, water immersion, and sand resistance.

Figure 6: Input a rough 4D GS model of a flower layered onto two backgrounds (a) and (b), harmonized into two different appearances with the same motion in (c) and (d).

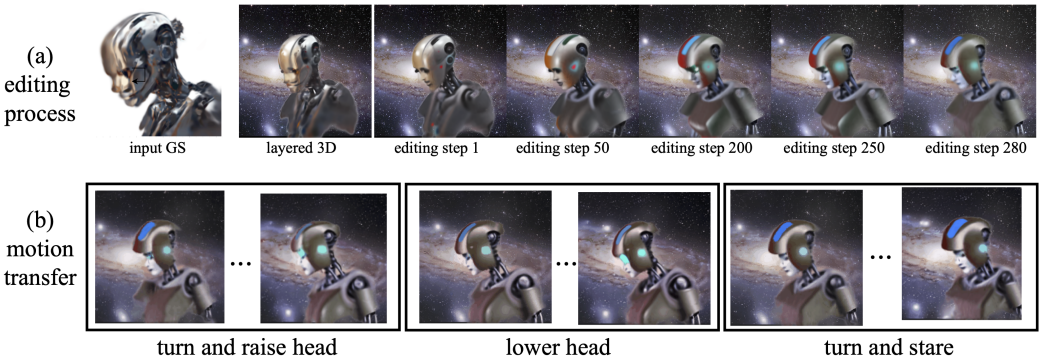

Figure 7: (a) shows input from a rough 3D GS model layered onto a universe background with visual results of the editing process; (b) demonstrates the edited 3D GS attached to different motion styles.

4.3 Baseline and Metrics

To evaluate the effectiveness of our novel 3D editing method for harmonization, we developed a quantitative metric. We used LucidDreamer and DreamGaussian as baselines due to their compatibility with Gaussian Splatting (GS) for background addition. Baselines involved adding backgrounds to rendered images using alpha masks from GS, followed by standard distillation scoring. We excluded methods not specifically designed for GS.

As shown in Fig. 8, we present edited images with backgrounds to test harmonization. We calculated the CLIP similarity [29] between image and text in the format "object + background". We tested

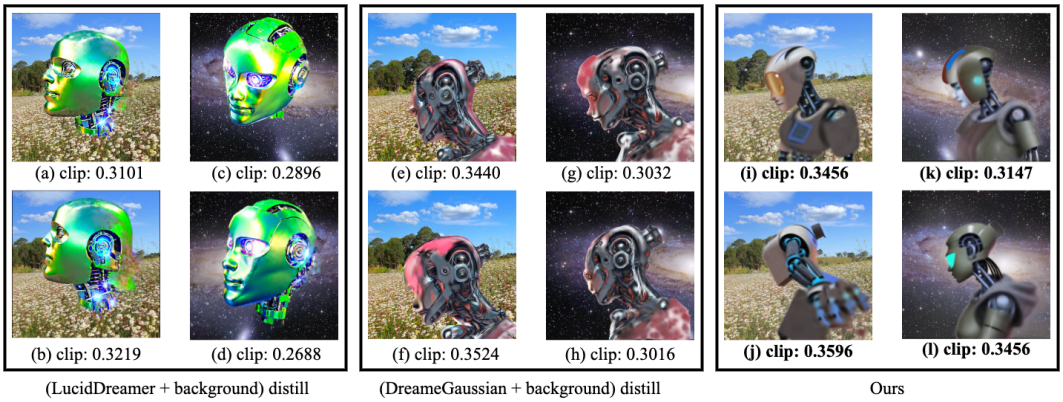

Figure 8: We layer universe and flower field backgrounds during the distillation process of LucidDreamer and DreamGaussian, presenting two randomly sampled views for each method and background. CLIP similarity scores for "robot in universe" and "robot in flower field" are shown below the images.

five categories (robot, cat, fish, cat astronaut) with five samples each on five backgrounds (moon, flower field, universe, sand, ocean), presenting the average CLIP scores in Table 1. And we call this metric CLIP-Harmonize Our method outperforms LucidDreamer, improving semantic alignment in backgrounds compared to DreamGaussian. For video quality, we sampled videos from the project pages of DreamMotion, Control4D, and DreamGaussian4D, which are recent works using distillation for video editing. Using the metrics CLIP[25], CLIP-Temp[25], and DOVER[45], we present the average scores of these samples in Table 2. CLIP measures text-frame average alignment, while CLIP temp assesses temporal semantic consistency by calculating the CLIP similarity between consecutive frames. Dover is a metric used to evaluate video quality.

Table 1: Average image harmonization CLIP score on 5 categories and 5 backgrounds

Method	LucidDreamer[23]	DreamGaussian[39]	Ours
CLIP-Harmonize↑	0.3270	0.3359	**0.3427**

Table 2: Average video quality with different metrics.

Method	DreamMotion[19]	Control4D[35]	DreamGaussian4D[33]	Ours
CLIP↑	0.4277	0.4584	0.3574	**0.3621**
CLIP-Temp↑	0.9900	**0.9967**	0.9872	0.9899
DOVER↑	0.505	0.5609	0.595	**0.6100**

4.4 Ablations

We conduct ablation studies to validate the effective ness of different components as shown in Fig 9. Our method allows us to get a better quality and harmonic 4D scene by adding both image and video inversions.

5 Limitations and Conclusions

The inversion method in diffusion can perform harmonization editing for 4D generation, allowing for 3D harmonization and 4D editing by decoupling appearances and motions. However, it is limited to short videos and restricted views due to base model constraints.

Figure 9: ablations on removing background, without video inverison, image inversion and full model

References

[1] S. Bahmani, X. Liu, Y. Wang, I. Skorokhodov, V. Rong, Z. Liu, X. Liu, J. J. Park, S. Tulyakov, G. Wetzstein, et al. Tc4d: Trajectory-conditioned text-to-4d generation. *arXiv preprint arXiv:2403.17920*, 2024.

[2] S. Bahmani, I. Skorokhodov, V. Rong, G. Wetzstein, L. Guibas, P. Wonka, S. Tulyakov, J. J. Park, A. Tagliasacchi, and D. B. Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *arXiv preprint arXiv:2311.17984*, 2023.

[3] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[4] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023.

[5] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh. Video generation models as world simulators. 2024.

[6] A. Cao and J. Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023.

[7] R. Chen, Y. Chen, N. Jiao, and K. 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*, pages 22246–22256, 2023.

[8] W.-H. Chu, L. Ke, and K. Fragkiadaki. Dreamscene4d: Dynamic multi-object scene generation from monocular videos. *arXiv preprint arXiv:2405.02280*, 2024.

[9] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

[10] B. Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

[11] D. Epstein, B. Poole, B. Mildenhall, A. A. Efros, and A. Holynski. Disentangled 3d scene generation with layout learning. *arXiv preprint arXiv:2402.16936*, 2024.

[12] Q. Gao, Q. Xu, Z. Cao, B. Mildenhall, W. Ma, L. Chen, D. Tang, and U. Neumann. Gaussianflow: Splatting gaussian dynamics for 4d content creation. *arXiv preprint arXiv:2403.12365*, 2024.

[13] A. Haque, M. Tancik, A. Efros, A. Holynski, and A. Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[14] Y. He, N. Murata, C.-H. Lai, Y. Takida, T. Uesaka, D. Kim, W.-H. Liao, Y. Mitsufuji, J. Z. Kolter, R. Salakhutdinov, et al. Manifold preserving guided diffusion. *arXiv preprint arXiv:2311.16424*, 2023.

[15] R. Henschel, L. Khachatryan, D. Hayrapetyan, H. Poghosyan, V. Tadevosyan, Z. Wang, S. Navasardyan, and H. Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text, 2024.

[16] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[17] 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*, pages 867–876, 2022.

[18] T. Jakab, R. Li, S. Wu, C. Rupprecht, and A. Vedaldi. Farm3d: Learning articulated 3d animals by distilling 2d diffusion. *arXiv preprint arXiv:2304.10535*, 2023.

[19] H. Jeong, J. Chang, G. Y. Park, and J. C. Ye. Dreammotion: Space-time self-similarity score distillation for zero-shot video editing. *arXiv preprint arXiv:2403.12002*, 2024.

[20] Y. Jiang, L. Zhang, J. Gao, W. Hu, and Y. Yao. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023.

[21] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

[22] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

[23] Y. Liang, X. Yang, J. Lin, H. Li, X. Xu, and Y. Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023.

[24] H. Ling, S. W. Kim, A. Torralba, S. Fidler, and K. Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. *arXiv preprint arXiv:2312.13763*, 2023.

[25] Y. Liu, X. Cun, X. Liu, X. Wang, Y. Zhang, H. Chen, Y. Liu, T. Zeng, R. Chan, and Y. Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023.

[26] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[27] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020.

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

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

[30] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[31] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation, 2021.

[32] D. Rempe, Z. Luo, X. Bin Peng, Y. Yuan, K. Kitani, K. Kreis, S. Fidler, and O. Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13756–13766, 2023.

[33] J. Ren, L. Pan, J. Tang, C. Zhang, A. Cao, G. Zeng, and Z. Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023.

[34] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[35] R. Shao, J. Sun, C. Peng, Z. Zheng, B. Zhou, H. Zhang, and Y. Liu. Control4d: Dynamic portrait editing by learning 4d gan from 2d diffusion-based editor. *arXiv preprint arXiv:2305.20082*, 2023.

[36] U. Singer, S. Sheynin, A. Polyak, O. Ashual, I. Makarov, F. Kokkinos, N. Goyal, A. Vedaldi, D. Parikh, J. Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023.

[37] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[38] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[39] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

[40] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation, 2024.

[41] P. Wang, Z. Fan, D. Xu, D. Wang, S. Mohan, F. Iandola, R. Ranjan, Y. Li, Q. Liu, Z. Wang, and V. Chandra. Steindreamer: Variance reduction for text-to-3d score distillation via stein identity, 2024.

[42] P. Wang, D. Xu, Z. Fan, D. Wang, S. Mohan, F. Iandola, R. Ranjan, Y. Li, Q. Liu, Z. Wang, et al. Taming mode collapse in score distillation for text-to-3d generation. *arXiv preprint arXiv:2401.00909*, 2023.

[43] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation, 2023.

[44] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and X. Wang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023.

[45] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023.

[46] D. Xu, H. Liang, N. P. Bhatt, H. Hu, H. Liang, K. N. Plataniotis, and Z. Wang. Comp4d: Llm-guided compositional 4d scene generation. *arXiv preprint arXiv:2403.16993*, 2024.

[47] Q. Yang, M. Feng, Z. Wu, S. Sun, W. Dong, Y. Wang, and A. Mian. Beyond skeletons: Integrative latent mapping for coherent 4d sequence generation, 2024.

[48] T. Yi, J. Fang, J. Wang, G. Wu, L. Xie, X. Zhang, W. Liu, Q. Tian, and X. Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models, 2024.

[49] Y. Yin, D. Xu, Z. Wang, Y. Zhao, and Y. Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023.

[50] Y. Yuan, V. Makoviychuk, Y. Guo, S. Fidler, X. Peng, and K. Fatahalian. Learning physically simulated tennis skills from broadcast videos. *ACM Trans. Graph*, 42(4), 2023.

[51] Y. Zhao, Z. Yan, E. Xie, L. Hong, Z. Li, and G. H. Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023.

[52] J. Zhu, P. Zhuang, and S. Koyejo. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance. In *The Twelfth International Conference on Learning Representations*, 2023.

A Appendix / supplemental material

A.1 Experimental Settings

We run the HarmonyDreamer pipeline on a single NVIDIA RTX A6000 GPU with 48G VRAM under Python 3.8 and CUDA 11.8 environment. Each harmonization editing step takes around 1 minute.

In our settings, we utilize an orbit camera with predefined and fixed intrinsic parameters. The extrinsic parameters of the camera can be parameterized by the azimuth angle ver, the horizon angle hor, and the radius r, thus $V = V(\text{ver}, \text{hor}, r)$. Additionally, the rendered image inherently includes the alpha mask of the assembly of Gaussian splats, which we denote as $G^t_{\mathbf{a}}(V)$.

A.2 More Qualitative Results

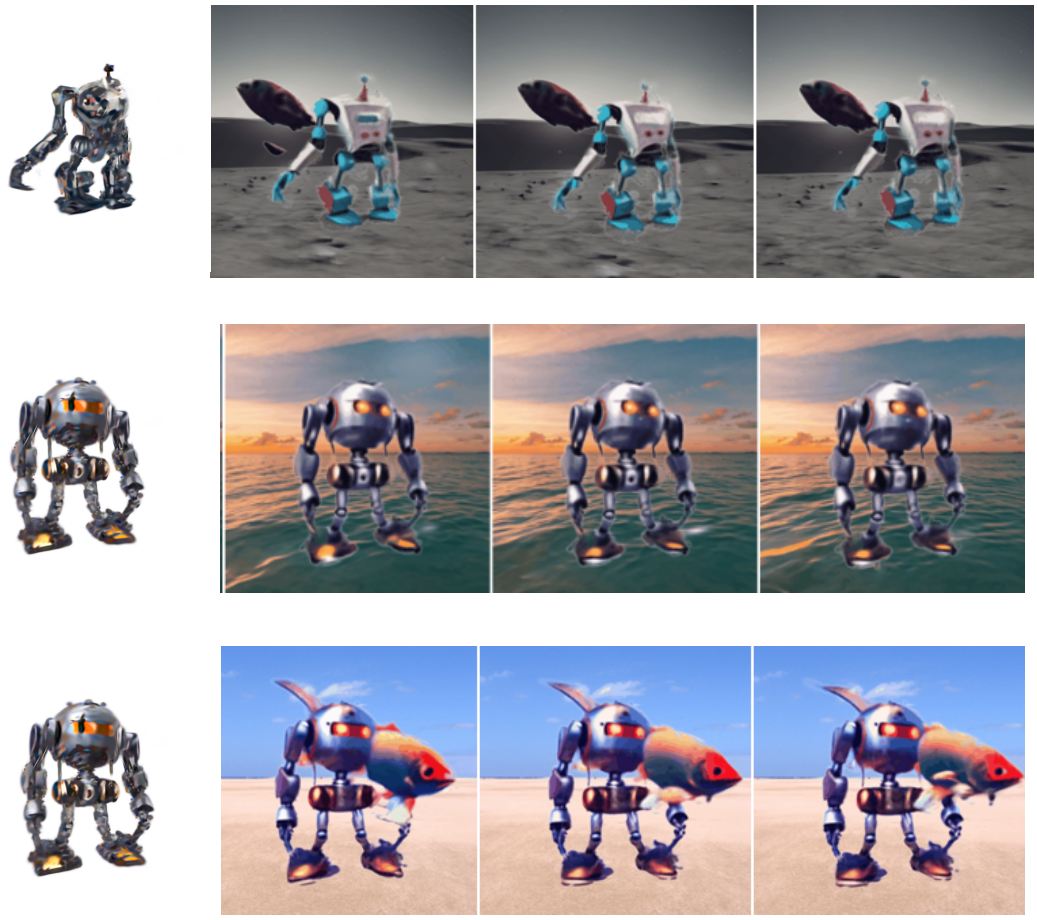

Figure 10: Emergent temporal semantics are observed after harmonization. For example, we see light reflection in the first row, water immersion in the second row, and sand resistance in the thrid row.

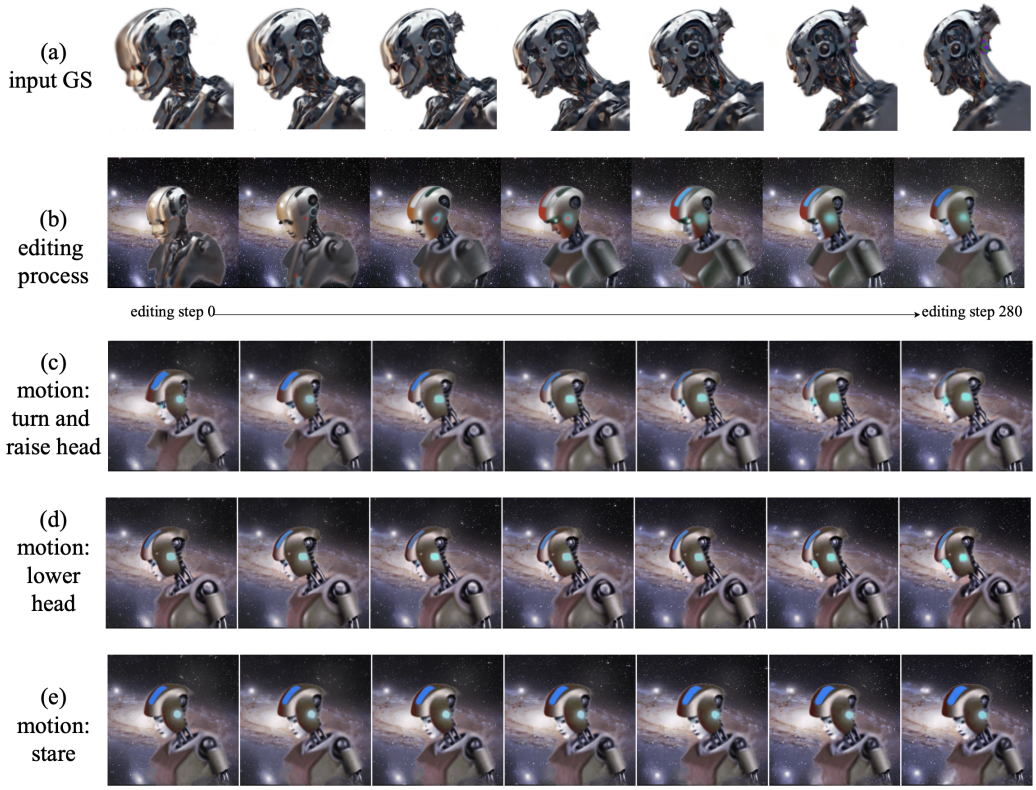

Figure 11: Expanding of Fig. 7(a) shows input from a rough 3D GS model layered onto a universe background with visual results of the editing process; (b) demonstrates the edited 3D GS attached to different motion styles.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claims are made based on the real experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitations of our method at the end of the main part.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our works does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We described our experimental results in the Experiments section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include our code in the supplementary material submission.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We described out experimental settings and details in the supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Our quantitative results are not accompanied by error bars, confidence intervals, or statistical significance tests.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We described out experimental settings and required resources in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work does not violate respects listed in the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Our work does not use photorealistic contents, which limits its potential for societal impact, as it avoids issues like disinformation, privacy violations, and generating fake profiles.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work does not pose risk of unsafe misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cited the source of the generated images/videos we used to conduct our work.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We include a anonymized URL to document all released new assets from our work.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.