Comp4D: Compositional 4D Scene Generation

Hanwen Liang[1] Dejia Xu[2] Neel P. Bhatt[2] Hezhen Hu[2] Hanxue Liang[3]
Konstantinos N. Plataniotis[1]

[1]University of Toronto [2]University of Texas at Austin [3]University of Cambridge

Abstract

The advancements in diffusion models for 2D and 3D content creation have sparked a surge of interest in generating 4D content. However, the scarcity of 3D scene datasets constrains current methodologies to primarily object-centric generation. To overcome this limitation, we present Comp4D, a novel framework for text-to-compositional 4D scene generation. Unlike conventional methods that generate a singular 4D representation of the entire scene, Comp4D innovatively employs a decompose-then-recompose strategy, constructing each 4D component within the scene separately. The framework first decomposes a textual input prompt into multiple object components and delineates their moving trajectories. After initializing the static 3D objects, we construct the compositional 4D scene by accurately positioning these objects along their designated paths. To refine the scene and motion, our method proposes a novel compositional score distillation technique involving trajectory-guided and object-centric sampling, utilizing pre-trained diffusion models across text-to-image, text-to-video, and text-to-3D domains for optimization. Extensive experiments demonstrate our superior 4D content creation capability compared to prior arts, showcasing superior visual quality, motion fidelity, and enhanced object interactions.

1. Introduction

The advancements in image and video diffusion models [25, 29, 31, 32] have significantly transformed generative AI by streamlining digital content creation. Traditional complex and expertise-dependent pipelines are being replaced by generative models capable of bringing sophisticated concepts to life from simple text prompts. This progress has also been extended to 3D generation, where score distillation techniques [22, 26, 34, 41, 42, 48] repurpose 2D diffusion models for 3D content synthesis. In parallel, video diffusion models are opening new directions for 4D generation. Recent methods typically rely on supervision from text prompts [2, 21, 35, 53, 56], images [19, 30, 53, 55], 3D models [53, 56], or monocular videos [11, 30, 53], to guide the generation process.

Despite these advancements, existing 4D generation methods are primarily object-centric and limited in the quality and realism of the synthesized motions. They often synthesize isolated objects (or multiple objects as a whole) constrained to limited motion within a confined region, overlooking global scene-level dynamics, including inter-object interactions and relative displacements. This limitation stems from the lack of comprehensive datasets that capture dynamic multi-object scenes. For example, Objaverse [7], widely used for 4D model training [22, 34], consists mostly of single objects at the world origin with minimal motion. As a result, current models struggle to generate scenes with realistic multi-entity dynamics. In this work, following [36], we define a "scene" as an environment composed of multiple objects, their spatial relationships, and dynamic interactions. To advance more realistic 4D scene generation, we extend the motion modeling from generating localized, object-centric motion of individual objects to modeling the global, scene-level dynamics and interactions between multiple objects in compositional environments.

To this end, we introduce **Comp4D**, the first framework for compositional 4D scene generation from text. Unlike prior methods that tightly synchronize object movements with the camera or focus only on local deformation, our framework models both global displacements between objects and local deformations of individual objects. To achieve this, we propose a novel *decompose-then-recompose* strategy that disentangles the compositional 4D scene generation into two stages: scene decomposition for constructing individual static 3D assets, and scene recomposition with motion modeling. In our framework, object motions are factorized into (1) *global displacements*, which describe inter-object positioning and movement across the scene, and (2) *local deformations*, which capture object-specific dynamics such as articulation or bending. We leverage a large language model (LLM) or manual annotations to generate kinematics-based trajectories that govern each object's global displacements. This factorization is

critical, as it alleviates the computational load on the deformation modules of 4D models, allowing them to concentrate on learning detailed local dynamics. Score distillation sampling (SDS) is utilized to optimize the motions with pre-trained image and video diffusion models. However, recomposing multiple moving objects in a shared 3D space introduces challenges such as frequent occlusions, which makes SDS-based training unstable [6, 41, 43]. To address this, we propose formulating each object as disjoint 3D Gaussians and introducing a novel *compositional score distillation* mechanism. This mechanism selectively renders either partial objects or the entire scene during motion optimization. Such a strategy proves to provide a powerful augmentation, improving the motion fidelity of each object, especially in scenarios where occlusions between objects are prevalent.

The generation of our 4D scene is conducted through the following steps. Given an input text description, we first leverage an LLM to decompose the scene by extracting entities and determining their attributes, such as scale. Static 3D objects are individually constructed using pre-trained 3D-aware diffusion models. Meanwhile, we can manually or take advantage of the LLM to design kinematics-based trajectory functions to guide the global displacement of objects. Subsequently, we re-compose the 4D scene with comprehensive motion learning. Each object's deformation field is optimized via the novel compositional score distillation, with objects moving along the pre-defined trajectories. Our key contributions can be summarized as follows.

- We introduce **Comp4D**, the first framework that achieves Compositional 4D scene generation from text prompts. By decomposing scenes into individual 3D objects and explicitly modeling their interactions, Comp4D moves beyond single-object generation to model global dynamics and interactions between multiple entities in a scene.
- We propose factorizing the motions of objects into global displacements and local deformations. Global displacement is guided by kinematics-based trajectories, either manually defined or generated by an LLM, allowing the 4D representation to focus on learning object-specific dynamics. The LLM facilitates effective and robust scene decomposition and trajectory design.
- Comp4D recomposes the scene by modeling motions with a novel compositional score distillation sampling, incorporating trajectory-guided and object-centric optimization. This design enables flexible switching between partial-object and whole-scene renderings, facilitating stable optimization of object appearance and motion even in scenarios involving occlusions.
- Extensive experiments demonstrate that Comp4D outperforms existing baselines in text alignment, visual quality, motion realism, and inter-object interaction, establishing a new standard for compositional 4D scene generation.

2. Related Works

2.1. 4D Content Creation

Text-guided diffusion models have advanced image and video synthesis, but limited large-scale 3D data hinders 3D generation. Score distillation sampling (SDS) [35] pioneers text-to-4D generation using NeRFs with HexPlane features. 4DFY [2] fuses image, video, and 3D-aware supervision for text-driven 4D synthesis. Consistent4D [11] addresses video-to-4D using RIFE [9] and super-resolution. AYG [21] introduces dynamic 3D Gaussians with a deformation field to disentangle motion from structure. [30, 53] further refine the quality of motion and texture. [15, 19, 54] generate 4D objects via multi-view consistent image diffusion. However, these works remain object-centric due to constraints in 3D-aware diffusion models. In contrast, our work is the first to tackle compositional 4D scene generation involving multiple interacting objects.

2.2. 4D Scene Representation

Building 4D scene representation allows for rendering novel views of dynamic objects. Recently, 3D Gaussian Splatting (3D-GS) [13] has shown advantages in both effectiveness and efficiency for 4D representation, leading to multiple directions to model temporal dynamics. Katsumata *et al.* [12] and 4DGS [44] define scales, positions, and rotations as functions of time while leaving other time-invariant properties of the static 3D Gaussians unchanged. Another direction involves directly extending 3D Gaussians to 4D with temporal slicing [8, 51]. There are also works leveraging a separate function to model the dynamic distribution of attributes' deformation for 3D Gaussians [18, 20, 47]. In this work, we adopt 3D Gaussians for our 3D content representation and use an additional Multi-layer Perceptron (MLP) to deform each set of 3D Gaussians. This disentangled 4D representation allows us to construct the static scene first and then focus on modeling the object's deformation.

2.3. Reasoning from Large Language Models

LLMs have emerged as a natural tool for real-world reasoning tasks [10, 16, 23, 28]. A popular approach to improving the reasoning capabilities of LLMs is to fine-tune models on domain-specific tasks [49]. Moreover, recent studies have explored techniques for incorporating multimodal information, such as images and videos, to enhance contextual understanding and improve the robustness of language models [33, 38]. Recently, LLMs have been used for generating trajectories in robotics applications. In [5, 14], dense trajectories were generated for a manipulator by an LLM in a zero-shot manner, demonstrating the potential of LLMs in trajectory generation. In this work, LLMs are used to generate trajectories of objects for 4D scene construction.

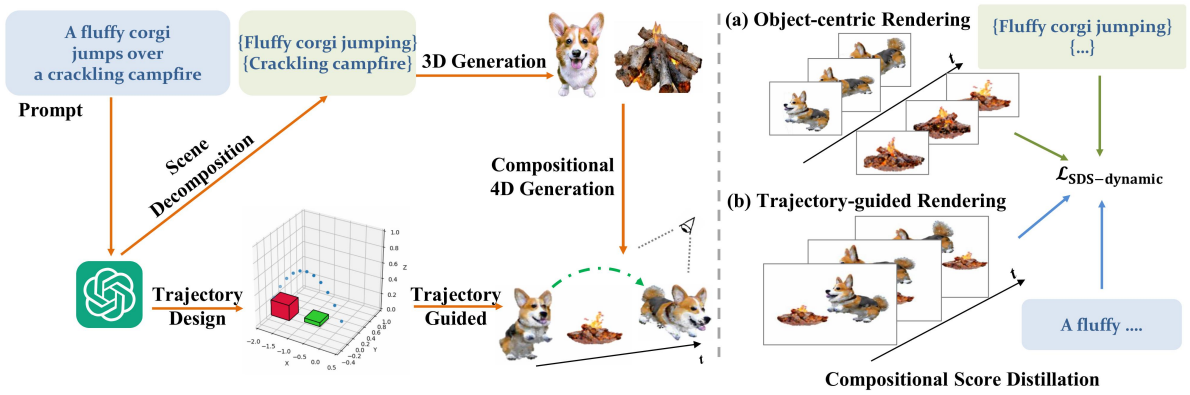

Figure 1. An overview of our proposed Comp4D. Given an input text description, we first perform scene decomposition and obtain multiple individual 3D components. We also design the object trajectories, which guide the global displacements of objects in a compositional 4D scene. Thanks to the Gaussian-based 4D representation, we propose a compositional score distillation that switches between object-centric rendering and trajectory-guided scene rendering flexibly at each training iteration.

3. Method

In this section, we illustrate the components of our proposed method in detail (Fig. 1). We start by introducing some preliminaries (Sec. 3.1) on 3D Gaussians and score distillation sampling. Then we introduce our decompose-then-recompose strategy and compositional 4D scene representation (Sec. 3.2). We later illustrate the compositional score distillation involving multiple diffusion models (Sec. 3.3). Finally, we discuss how we leverage LLMs for scene decomposition including scale assignment and trajectory design (Sec. 3.4).

3.1. Preliminaries

3D Gaussian Splatting 3D Gaussian Splatting (3D-GS) [13] parameterizes a 3D scene as a set of 3D Gaussians. Each Gaussian is defined with a center position μ, covariance Σ, opacity α, and color c modeled by spherical harmonics. Unlike implicit representation methods such as NeRF [24], which renders images based on volumetric rendering, 3D-GS renders images through a tile-based rasterization operation and achieves real-time rendering speed. Starting from a set of points randomly initialized in the unit sphere, each point is designated a 3D Gaussian, which can be queried as follows:

$$G(x) = e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}, \quad (1)$$

where x is an arbitrary position in the 3D scene. During the rendering process, the 3D Gaussians $G(x)$ are first transformed to 2D Gaussians $G^{'}(x)$ on the image plane. Then a tile-based rasterizer is designed to efficiently sort the 2D Gaussians and employ α-blending:

$$C(r) = \sum_{i \in N} c_i \sigma_i \prod_{j=1}^{i-1}(1-\sigma_j), \quad \sigma_i = \alpha_i G'(r), \quad (2)$$

where r is the queried pixel position, N denotes the number of sorted 2D Gaussians associated with the queried pixel, c_i and α_i denote the color and opacity of the i-th Gaussian. In our experiments, we empirically simplify the color of Gaussians to diffuse color for the sake of efficient training.

Score Distillation Sampling Current methodologies for text-to-3D or 4D generation typically involve iterative optimization of a scene representation with supervisory signals from pre-trained diffusion models [26, 42]. Initially, rendering of the 3D or 4D scene is acquired in the form of an image or sequence of images. Random noise is added to the rendered images, and a pre-trained diffusion model is employed to de-noise the images. The estimated gradient from this process is utilized to update the 3D or 4D representations. Specifically, employing a 3D representation parameterized by θ and a rendering method g, the rendered images are generated as $x = g(\theta)$. To align the rendered image x with samples obtained from the diffusion model ϕ, the diffusion model employs a score function $\hat{\epsilon}_\phi(x_t; y, t)$ to predict a noise map $\hat{\epsilon}$, given the noise level t, noisy input x_t and text embeddings y. By evaluating the difference between the Gaussian noise ϵ added to the rendered images x and the predicted noise $\hat{\epsilon}$, this score function updates the parameter θ with gradient formulated as:

$$\nabla_\theta \mathcal{L}_{SDS}(\phi, x = g(\theta)) = w(t)(\hat{\epsilon}_\phi(x_t; y, t) - \epsilon)\frac{\partial x}{\partial \theta}, \quad (3)$$

where $w(t)$ is a weighting function. Using SDS for 4D generation requires coordinated guidance to achieve realistic outcomes in terms of appearance, 3D structure, and motion [2]. This often involves the utilization of hybrid SDS, which combines both image-based and video-based diffusion models [21]. For our compositional 4D scene generation task, we develop a compositional SDS technique that is applied to a varying number of assets in the scene.

3.2. Compositional 4D Representation

We develop a decompose-then-recompose strategy to build compositional 4D scenes. Given a text description, we first decompose the description into multiple assets that make up the scene. Each asset is assigned a scale and a moving trajectory, either manually or through LLM models. The 4D scene is then constructed by recomposing these individual objects. In Fig. 1, we use two objects for illustration. Our framework is easily applicable to more objects.

For each object, we utilize a set of static 3D Gaussians along with an MLP-based deformation network. The MLP network takes in (x, y, z, t) coordinates as input and outputs the 3D deformation of point locations. Following previous works [24, 37], the input coordinates are processed with positional encoding as a 32-dimensional vector to enable high-frequency feature learning. This architecture design supports decoupled learning of the static attributes of an object (*e.g.* geometry and texture) and the local motion information. We start our training stage by optimizing the static 3D Gaussian attributes. Once they converge, we introduce the deformation field and freeze partial 3D Gaussian attributes (i.e. covariance, opacity, and color) to stabilize the training process. However, naively optimizing the deformation field leads to unpleasant results. This is primarily because the MLP modulates each point location individually, ignoring the overall rigidity of the object. Similar to AYG [21], we adopt rigidity constraints to ensure that the deformation of each Gaussian is consistent with its k-nearest neighbors,

$$\mathcal{L}_{\text{rigidity}}(x) = \frac{1}{k}\sum_{i=1}^{k} ||\Delta_x - \Delta_{x_{NN_i}}||. \tag{4}$$

Moreover, to avoid flickering motion, we introduce additional regularization loss components that penalize sudden changes in the acceleration of each 3D Gaussian,

$$\mathcal{L}_{\text{acc}}(x, t) = ||\Delta_{x,t} + \Delta_{x,t+2} - 2\Delta_{x,t+1}||. \tag{5}$$

For the whole scene optimization, thanks to the explicit nature of 3D Gaussians, at rendering time, we can selectively render a single object or multiple objects, and perform compositional SDS. This enables direct and better supervision over the motion learning of each object as well as their interactions. Meanwhile, since the objects are separately represented as 3D Gaussians, we need explicit constraints to prevent the objects from intersecting with each other. If the objects have overlapping parts, the rendered image will show collapsed shapes, resulting in unstable gradients from score distillation. To this end, we draw inspiration from CG3D [39] to incorporate a physics-based contact loss that avoids the collision of multiple objects. For one object, we ensure the contact angle θ_j for each 3D Gaussian with mean μ_j to be acute:

$$\begin{aligned} \theta_j &= (cr - \mu_i) \cdot (\mu_j - \mu_i), \\ \mathcal{L}_{\text{contact}} &= -\theta_j[\theta_j < 0], \end{aligned} \tag{6}$$

where μ_i denotes the mean of Gaussians from another object closest to μ_j, and cr is the center of the current object.

3.3. 4D Scene Optimization via Compositional Score Distillation

We start the 4D scene generation by constructing each static 3D component. In the subsequent whole scene optimization, we propose compositional SDS, which involves trajectory-guided scene optimization and object-centric motion learning. We illustrate these parts in detail in the following sections.

Static 3D Object Construction To ensure both photo-realism of texture and consistent geometry, we draw inspiration from Magic123 [27] and 4DFY [2] to incorporate the joint distillations of an image diffusion [31] and a 3D-aware diffusion model [34]. Specifically, we adopt the weighted combination of two sets of score distillation losses. Given a batch of rendered image x and text embeddings y, the loss function is formulated as follows,

$$\mathcal{L}_{\text{static}}(x, y) = \omega_1(\epsilon_{\text{sd}}(x_t; y, t) - \epsilon_1) + \omega_2(\epsilon_{\text{mv}}(x_t; y, t) - \epsilon_2) \tag{7}$$

where ω_1 and ω_2 are coefficients for the score distillation loss of Stable Diffusion [31] and MVDream [34].

Trajectory-Guided Scene Optimization After the initial construction of static 3D assets, we focus on the object's motion learning. At the scene level, the object's motion can be decomposed into global displacement and local deformation. The global displacement, represented by the moving trajectory, can be designed manually or by LLMs. We sample uniformly from the trajectory function, $F(\cdot)$, and obtain the object locations at arbitrary timesteps t_i. Objects are rotated accordingly such that their canonical orientation faces toward the next location along the trajectory $\overrightarrow{R_i} = (F(t_{i+1}) - F(t_i))$. Thanks to MVDream [34], which generates objects in their canonical orientation, our static stage produces objects facing the same direction (*e.g.* $\overrightarrow{R_0} = (1, 0, 0)$), ensuring that our rotation strategy will produce objects moving towards their head direction. Given normalized head direction $A = \frac{\overrightarrow{R_0}}{||\overrightarrow{R_0}||}$ and $B = \frac{\overrightarrow{R_i}}{\overrightarrow{R_i}}$, the axis of rotation v is obtained as $v = A \times B$. The angle of rotation θ is determined by $\cos(\theta) = A \cdot B$. We then obtain the skew-symmetric matrix $\mathbf{K}$ as follows,

$$\mathbf{K} = \begin{bmatrix} 0 & -v_z & v_y \\ v_z & 0 & -v_x \\ -v_y & v_x & 0 \end{bmatrix}, \tag{8}$$

which is then used in Rodrigues' rotation formula to obtain the final rotation matrix $\mathbf{R}$,

$$\mathbf{R} = \mathbf{I} + (\sin\theta)\mathbf{K} + (1 - \cos\theta)\mathbf{K}^2. \tag{9}$$

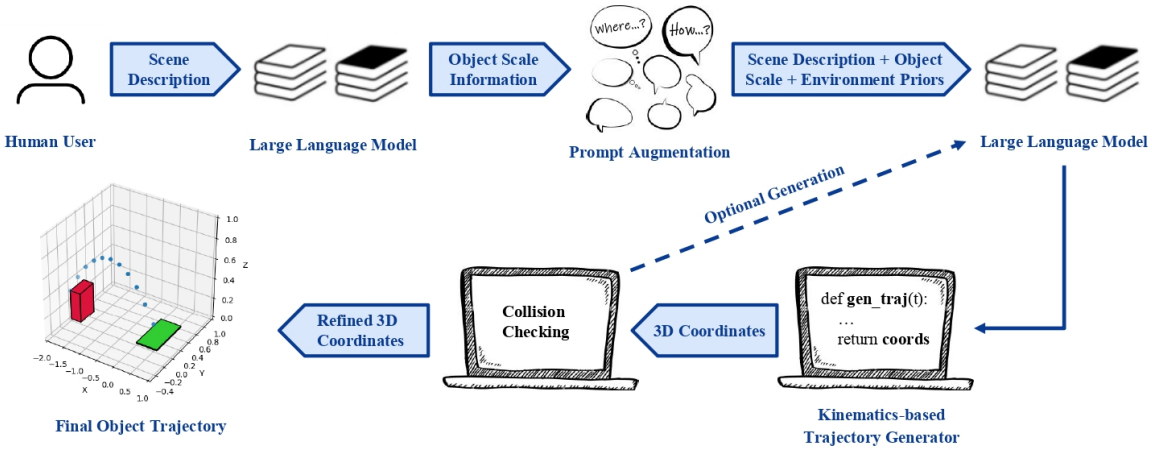

Figure 2. The pipeline for scene decomposition and trajectory design with LLMs. First, a scene description is provided by a human user as a prompt to an LLM which yields the object components as well as the relative object scales. Subsequently, the LLM is prompted with environmental constraints to return a trajectory function, which takes timestep as an input and returns the corresponding object's 3D positions. After the collection of a set of positions, collision checking is performed to truncate the trajectory if any collision occurs. Optionally, premature collisions can be mitigated by re-querying the LLMs for an improved trajectory function.

Thanks to the predefined trajectory, our framework supports distilling objects with long-range motion and multi-concept interactions, which is difficult to achieve using previous baselines.

Besides global displacement, we utilize a deformation MLP for each set of 3D Gaussian for local motion learning. To better learn the deformation field, we leverage a text-to-video diffusion model [1] to formulate the score distillation loss. Similar to distilling a static 3D object via an image diffusion model, score distillation via a video diffusion model ensures that the renderings at consecutive frames form a natural video aligned with the text prompt. As observed in previous works [2, 21], image diffusion models usually generate a more realistic appearance compared to video diffusion models. Therefore, we jointly distill the score from image diffusion on individual frames to ensure texture quality. The loss function can be formulated as follows,

$$\mathcal{L}_{\text{traj}}(x, y) = \omega_{\text{img}}(\epsilon_{\text{sd}}(x_t; y, t) - \epsilon_1) + \omega_{\text{vid}}(\epsilon_{\text{vid}}(x_t; y, t) - \epsilon_2), \tag{10}$$

where x is the generated image sequence of the whole scene and y is the text prompt. ω_{img} and ω_{vid} are coefficients for the score distillation loss from image- and video-based diffusion models.

Object-Centric Motion Learning The compositional design in our framework enables arbitrary combinations of objects during rendering. At each training iteration, we flexibly choose to render either the entire scene or selected subsets of objects. This provides great flexibility in rendering scenes with diverse appearances. Such diversity provides rich augmentations that are crucial for the stable optimization of score distillation loss, particularly in mitigating the negative impact of object occlusions on motion supervision. When rendering partial objects for motion learning, we modify the text prompt by removing references to inactive entities, ensuring focused learning of the active object's deformations. Following the whole scene optimization, we supervise the object-centric motion learning with joint score distillation losses (Eq. 10).

3.4. LLM Guided Scene Decomposition and Trajectory Design

We leverage large language models (LLMs) to automate the processes of scene decomposition and trajectory design, thereby offloading complexity from the 4D representation and allowing score distillation models to focus on learning realistic local deformations. The overall pipeline is illustrated in Fig. 2.

Scene Decomposition Given a text prompt, we first use an LLM to decompose the scene into multiple components and give distinct descriptions of each component. These descriptions are directly used as prompts in object-centric motion learning. Since most 3D-aware diffusion models are trained on normalized, unit-scale synthetic assets, inferring the correct relative scale for each object becomes crucial for a realistic and reasonable composition of the scene. Recent studies [4, 17] show that LLM (e.g. GPT-4) demonstrates a remarkable ability to reason with common sense knowledge. Therefore, LLM is further prompted to make reasonable assumptions of the relative scale of the objects. The inferred relative scales are then used to adaptively rescale each 3D object.

Trajectory Design Through Kinematics Templates We further leverage the reasoning capability of LLM to select physics-based formulas to govern the displacement of objects. To streamline and simplify the task, we instruct the model to assume that one reference object is always positioned at the camera origin and solely design the trajectory of relative displacement between the objects in the coordinate system relative to the reference object. The trajectory follows kinematics-based equations such as uniform linear motion and parabolic motion. The LLM also adeptly determines the initial positions and velocities based on the scene semantics, ensuring that trajectories are contextually consistent and physically plausible. While we take advantage of LLM to obtain trajectories, our compositional 4D scene generation framework also supports optional manual editing or custom-designed trajectories to accommodate specific scene requirements.

	Human Preference↑					*QAlign Metrics*↑				*CLIP* ↑	*Efficiency*↑
Method	3DC	AQ	MF	TA	Overall	Img-quality	Img-aesthetic	Vid-quality	Vid-aesthetic	Img-Text	Rendering fps
4DFY [2]	34%	28%	26%	34%	30%	2.031	1.767	2.465	1.973	28.76	4
Animate124 [55]	26%	24%	20%	28%	25%	1.434	1.484	1.948	1.654	24.16	4
Ours†	**40%**	**48%**	**54%**	**38%**	**45%**	**2.655**	**2.057**	**2.912**	**2.187**	**30.54**	–
TC4D [3]	35%	**52%**	30%	40%	40%	2.617	1.936	2.838	2.109	29.66	4
Ours	**65%**	48%	**70%**	**60%**	**60%**	**2.931**	**2.190**	**3.367**	**2.461**	**31.98**	**70**

Table 1. Quantitative comparison between our method and baselines. Human preference evaluations include 3D geometry consistency (3DC), appearance quality (AQ), motion fidelity (MF), text alignment (TA), and overall score. QAlign metrics assess quality and aesthetics of rendered images/videos, and CLIP measures text–image alignment. "Ours†" denotes down-sampling our original generations to match the low-resolution settings of 4DFY [2] and Animate124 [55].

Optional Trajectory Refinement via Collision Checking Despite curated prompt engineering, some LLM-generated trajectories may cause unintended collisions. To mitigate this, we incorporate an optional trajectory refinement step to check for physical feasibility. A sequence of points along the trajectory is sampled as objects' centers at corresponding timestamps. To efficiently simulate object occupancy, each object is approximated with a pre-sized rectangular cuboid. They are also rotated so that the canonical orientation of objects faces the next sampled location. Despite its simplicity, this strategy works very well in practice for coarse-level spatial planning. After obtaining the object placement at each timestamp, Eq. 6 is utilized for collision checking. Trajectories are automatically truncated at the first detected collision to ensure stable rendering. If the truncated path is too short to enable meaningful motion, the system re-prompts the LLM to regenerate a new trajectory. Together, these steps form a robust and automated scene planning pipeline that produces valid, scalable, and diverse scene configurations with minimal manual intervention.

4. Experiments

4.1. Implementation Details

Given a text prompt, we use GPT-4 to decompose the scene into multiple components, assigning appropriate scales and designing motion trajectories for each asset. For asset generation, inspired by [2, 27], we first create static 3D objects using joint score distillation from MVDream [34] and Stable Diffusion 2.1 [31]. These objects are then converted into point clouds to initialize 3D Gaussians. Each object is represented with 60,000 Gaussian points. In the compositional optimization stage, we randomly assign training iterations to adopt single-object rendering (with a probability of 0.2) or whole-scene rendering (with a probability of 0.8). In each iteration, we render 16 frames via uniformly sampled timesteps. We use the frozen diffusion models, Zeroscope [1] and Stable Diffusion 2.1 [31], to provide SDS supervision. We compare our method with two prior object-centric text-to-4D generation works, 4DFY [2] and Animate124 [55], and a trajectory-based concurrent work, TC4D [3].

4.2. Main Results

Quantitative Comparison We evaluated our method against baselines using 20 text prompts describing diverse compositional scenes with 2-4 assets. First, a user study is performed involving 30 participants from diverse backgrounds. Participants evaluated rendered videos of 4D scenes based on four key properties, following the practice in [2, 3]: 3D Geometry Consistency (3DC), Appearance Quality (AQ), Motion Fidelity (MF), and Text Alignment (TA). For each method, we demonstrate four views (0°,90°,180°, 270°) videos for preference selection. We report the percentage of user preferences overall and for each property. In the absence of ground truth for unsupervised text-to-4D scene generation, we employed non-reference quality-assessment models for images and videos. Q-Align [45] is a recently proposed large multi-modal model fine-tuned from mPLUG-Owl2 [52] using in-the-wild image and video quality assessment datasets. It provides quality assessment functionality for images and videos in terms of aesthetics and quality, achieving state-of-the-art performance in alignment with human ratings on existing quality assessment benchmarks. We report the average scores on four views (0°,90°,180°, 270°) of our test samples. We also use CLIP [40] to measure the text-image semantic alignments. The results are reported in Tab. 1. It can be observed that our method outperforms baselines in most metrics by a large margin.

Qualitative Comparison In Fig. 3, we provide a detailed visualization of generated scenes with multiple assets at different timestamps from various views. For 4DFY [2] and Animate124 [55], we show the scenes from one view at timestamps of 0, 0.2, and 1s. For TC4D [3] and our method, we show views at uniformly sampled timestamps from 0 to 1s. As shown in the image, our framework excels in generating lifelike objects with expansive motions while enhancing fidelity in object interactions. As indicated by the yellow contours in Fig. 3, we can observe the changes in body shape as the frog jumps, the distinct flapping of butterfly wings, and variations in body contours as the dog runs. Also, the objects move following the pre-defined trajectory and display more frequent and realistic interactions. As indicated by the yellow circles, two fish swim around a rock showing large displacements, the frog stretches out its legs on the lotus leaf, the butterfly settles on the petal, and the dog steps on a skateboard. Comparatively, 4DFY and Animate124 exhibit limited movements where the objects are fixed at the origin. Though TC4D shows object movements guided by trajectory, the local deformations are imperceptible, with texture flickering to simulate local motions.

Resolution and Speed Due to NeRF [24] expensive rendering cost, 4DFY [2] uses resolution of 160×288, and Animate124 [55] uses 80×144 in score distillation. In contrast, our method can ren-

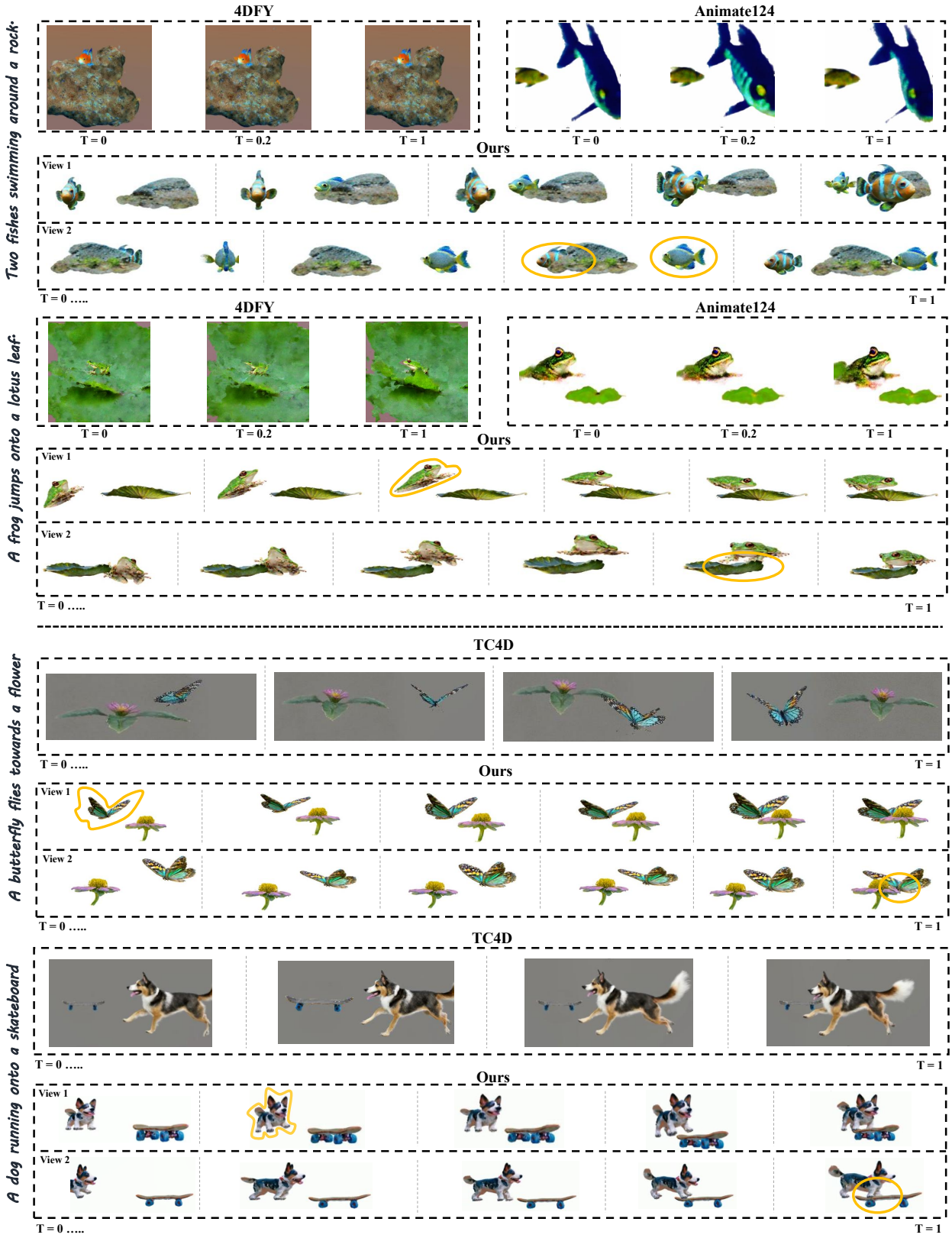

Figure 3. Comparison with object-centric 4D generation baselines (4DFY and Animate124) and trajectory-based work (TC4D). Comp4D generates more realistic object motions and interactions, with large global displacements and noticeable local deformations. The baselines exhibit limited movements with objects fixed at the origin (example 1, 2), or show imperceptible local deformations (example 3, 4).

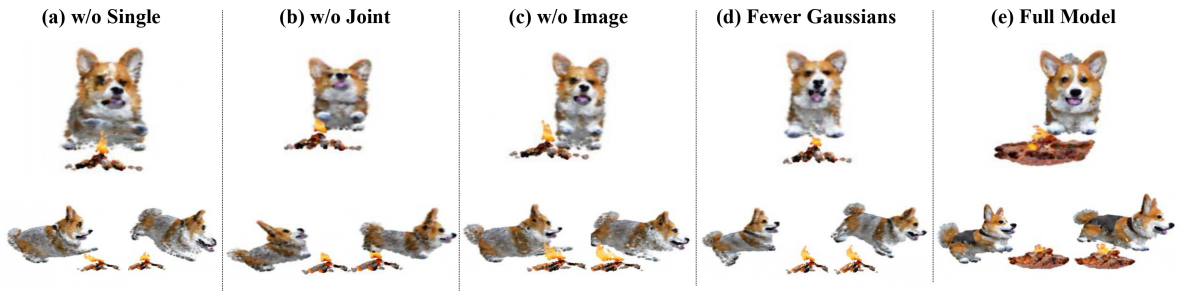

Figure 4. Ablation studies on the proposed components. The first row shows the front view. The second row shows the side view. To save computation costs, (a)-(d) are conducted using same fewer number of Gaussians.

Succeed at:	1st trial	2nd-3rd trial	4th-6th trial	avg # of trials
2-object-straight	100%	-	-	1.0
2-object-curved	80%	100%	-	1.3
3-object-straight	70%	90%	100%	1.6
3-object-curved	40%	70%	100%	2.5
4-object-straight	30%	70%	100%	3.2
4-object-curved	20%	40%	100%	4.4

Table 2. Cumulative success rate of LLM-generated trajectory in different settings, i.e. number of objects (two, three, and four) and trajectory types (straight and curved).

Settings	QAlign-Img-quality↑	QAlign-Img-aesthetic↑	QAlign-Vid-quality↑	QAlign-Vid-aesthetic↑
w/o Single	1.8252	1.6455	2.4082	1.9062
w/o Joint	1.9893	1.8789	2.4102	1.9512
w/o Image	1.8613	1.7715	2.3926	1.9014
Fewer GS	1.9131	1.8301	2.7285	2.0039
Full	**2.4785**	**1.9004**	**2.9023**	**2.1621**

Table 3. Ablation studies on our proposed components. We employ QAlign metrics, including quality and aesthetic evaluations on both rendered images and videos.

der video at a resolution of 320×576. To have a fair comparison, we down-sampled our generation to 160×288 when comparing with them. For TC4D, we adopt the same higher resolution. About training cost, with a single NVIDIA H100, 4DFY takes around 30h for three-stage optimization, Animate124 takes around 25h, and TC4D takes around 32h. In our method, obtaining 3D assets takes 12h, and 4D motion optimization takes 8h. At inference time, thanks to the efficient Gaussian representation, with a single NVIDIA A100, our 4D scene representation renders at 70 FPS at 320×576 resolution, much more efficient than the baselines.

Robustness of LLM model in trajectory design Our framework integrates collision checking, trajectory truncation, and re-generation to improve LLM-based trajectory generation. We evaluate success rates on 10 two-object, 10 three-object, and 10 four-object scenes. For each scene, GPT-4 is prompted to generate one straight and one curved path. A trajectory is considered successful if it avoids collisions. The cumulative success rate for each case w.r.t. the number of trials is shown in Tab. 2. For simpler scenarios with two objects moving along straight paths, the framework achieves a 100% success rate on the first trial. For more complex cases, such as three objects with curved trajectories, the success rate is 40% on the first trial, requiring an average of 2.5 trials to generate valid paths. In the most challenging scenario involving four objects with curved paths, an average of 4.4 trials is needed for successful trajectory generation. Overall, the LLM demonstrates strong performance in generating collision-free trajectories, even for complex multi-object scenarios. At this point, we emphasize that our method is flexible and can incorporate human involvement for the initial trajectory design.

4.3. Ablation Studies

We evaluate the effectiveness of our components in Fig. 4 and Tab. 3. To save computation costs, we utilize 3D Gaussians containing 20,000 points to represent each object. In Fig. 4(a), removing object-centric rendering ("w/o Single") results in poor geometry due to occlusion during optimization. Fig. 4(b) shows that disabling joint rendering ("w/o Joint") reduces motion realism and inter-object interactions. Fig. 4(c) illustrates that excluding the image diffusion SDS loss yields weaker textures compared to (d), where it is included. In Fig. 4(d), the model training and losses are kept the same as the full model (e), except that the number of 3D Gaussians we generate in the static stage is fewer. The figure also shows that using fewer Gaussians results in less detailed texture and less realistic geometry. In summary, using full model (e) delivers the best qualitatively results. Quantitatively, Tab. 3 confirms that removing either object-centric or joint rendering noticeably degrades image and video quality, while the full model provides the best results.

5. Conclusion

In this work, we present Comp4D, a novel framework for generating compositional 4D scenes from text input. Given a compositional scene description, we first leverage GPT-4 to generate object prompts for the independent creation of 3D objects. Subsequently, it is tasked to design the trajectory for the moving objects. This predefined trajectory then guides the compositional score distillation process, which optimizes a composable 4D representation comprising deformable 3D Gaussians for each object. Our experiments demonstrate that Comp4D significantly surpasses existing text-to-4D generation methods in terms of visual quality, motion fidelity, and object interactions.

References

[1] Zeroscope. https://huggingface.co/cerspense/zeroscope_v2_576w, 2023. 5, 6, 12

[2] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *arXiv preprint arXiv:2311.17984*, 2023. 1, 2, 3, 4, 5, 6, 12, 13

[3] Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. Tc4d: Trajectory-conditioned text-to-4d generation. In *European Conference on Computer Vision*, pages 53–72. Springer, 2024. 6, 13

[4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 5

[5] Arthur Bucker, Luis Figueredo, Sami Haddadin, Ashish Kapoor, Shuang Ma, Sai Vemprala, and Rogerio Bonatti. Latte: Language trajectory transformer. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7287–7294. IEEE, 2023. 2

[6] Yongwei Chen, Tengfei Wang, Tong Wu, Xingang Pan, Kui Jia, and Ziwei Liu. Comboverse: Compositional 3d assets creation using spatially-aware diffusion guidance. In *European Conference on Computer Vision*, pages 128–146. Springer, 2024. 2

[7] 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*, pages 13142–13153, 2023. 1

[8] Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 4d gaussian splatting: Towards efficient novel view synthesis for dynamic scenes. *arXiv preprint arXiv:2402.03307*, 2024. 2

[9] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision*, pages 624–642. Springer, 2022. 2

[10] Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pages 287–318, Auckland, New Zealand, 2022. PLMR. 2

[11] Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023. 1, 2

[12] Kai Katsumata, Duc Minh Vo, and Hideki Nakayama. An efficient 3d gaussian representation for monocular/multi-view dynamic scenes. *arXiv*, 2023. 2

[13] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 2, 3

[14] Teyun Kwon, Norman Di Palo, and Edward Johns. Language models as zero-shot trajectory generators. In *2nd Workshop on Language and Robot Learning: Language as Grounding*, 2023. 2

[15] Bing Li, Cheng Zheng, Wenxuan Zhu, Jinjie Mai, Biao Zhang, Peter Wonka, and Bernard Ghanem. Vivid-zoo: Multi-view video generation with diffusion model. *Advances in Neural Information Processing Systems*, 37:62189–62222, 2024. 2

[16] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Conference on Computer Vision and Pattern Recognition*, pages 10955–10965, New Orleans, LA, USA, 2022. IEEE. 2

[17] Yunxin Li, Longyue Wang, Baotian Hu, Xinyu Chen, Wanqi Zhong, Chenyang Lyu, and Min Zhang. A comprehensive evaluation of gpt-4v on knowledge-intensive visual question answering. *arXiv preprint arXiv:2311.07536*, 2023. 5

[18] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. *arXiv preprint arXiv:2312.16812*, 2023. 2

[19] Hanwen Liang, Yuyang Yin, Dejia Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. *arXiv preprint arXiv:2405.16645*, 2024. 1, 2

[20] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. *arXiv preprint arXiv:2312.03431*, 2023. 2

[21] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. *arXiv preprint arXiv:2312.13763*, 2023. 1, 2, 3, 4, 5, 13

[22] 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. *arXiv preprint arXiv:2303.11328*, 2023. 1

[23] Yujie Lu, Pan Lu, Zhiyu Chen, Wanrong Zhu, Xin Eric Wang, and William Yang Wang. Multimodal procedural planning via dual text-image prompting, 2023. 2

[24] 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. *Communications of the ACM*, 65(1):99–106, 2021. 3, 4, 6

[25] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

[26] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 3

[27] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 4, 6, 12

[28] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*, 2019. 2

[29] 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. 1

[30] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 1, 2

[31] 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*, pages 10684–10695, 2022. 1, 4, 6, 12

[32] 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*, 2022. 1

[33] Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S. Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. *arXiv preprint arXiv:2309.16534*, 2023. 2

[34] 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. 1, 4, 6, 12

[35] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 1, 2

[36] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 1

[37] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 4

[38] Archana Tikayat Ray, Anirudh Prabhakara Bhat, Ryan T. White, Van Minh Nguyen, Olivia J. Pinon Fischer, and Dimitri N. Mavris. Examining the potential of generative language models for aviation safety analysis: Case study and insights using the aviation safety reporting system (asrs). *Aerospace*, 10(9), 2023. 2

[39] Alexander Vilesov, Pradyumna Chari, and Achuta Kadambi. Cg3d: Compositional generation for text-to-3d via gaussian splatting. *arXiv preprint arXiv:2311.17907*, 2023. 4

[40] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022. 6

[41] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 1, 2

[42] 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. *arXiv preprint arXiv:2305.16213*, 2023. 1, 3

[43] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[44] 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. 2

[45] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 6

[46] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22227–22238, 2024. 12

[47] Ke Xing, Hanwen Liang, Dejia Xu, Yuyang Yin, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. Tip4gen: Text to immersive panorama 4d scene generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9267–9276, 2025. 2

[48] 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 {\deg} views. *arXiv preprint arXiv:2211.16431*, 2022. 1

[49] Yunhao Yang, Neel P Bhatt, Tyler Ingebrand, William Ward, Steven Carr, Zhangyang Wang, and Ufuk Topcu. Fine-tuning language models using formal methods feedback. *arXiv preprint arXiv:2310.18239*, 2023. 2

[50] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023. 12

[51] Zeyu Yang, Hongye Yang, Zijie Pan, Xiatian Zhu, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*, 2023. 2

[52] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023. 6

[53] Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023. 1, 2

[54] Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 4diffusion: Multi-view video diffusion model for 4d generation. *Advances in Neural Information Processing Systems*, 37:15272–15295, 2024. 2

[55] 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. 1, 6, 12

[56] Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text-and image-guided 4d scene generation. *arXiv preprint arXiv:2311.16854*, 2023. 1