# Splat4D: Diffusion-Enhanced 4D Gaussian Splatting for Temporally and Spatially Consistent Content Creation

MINGHAO YIN, The University of Hong Kong, Hong Kong
YUKANG CAO, Nanyang Technological University, Singapore
SONGYOU PENG, Google DeepMind, USA
KAI HAN*, The University of Hong Kong, Hong Kong

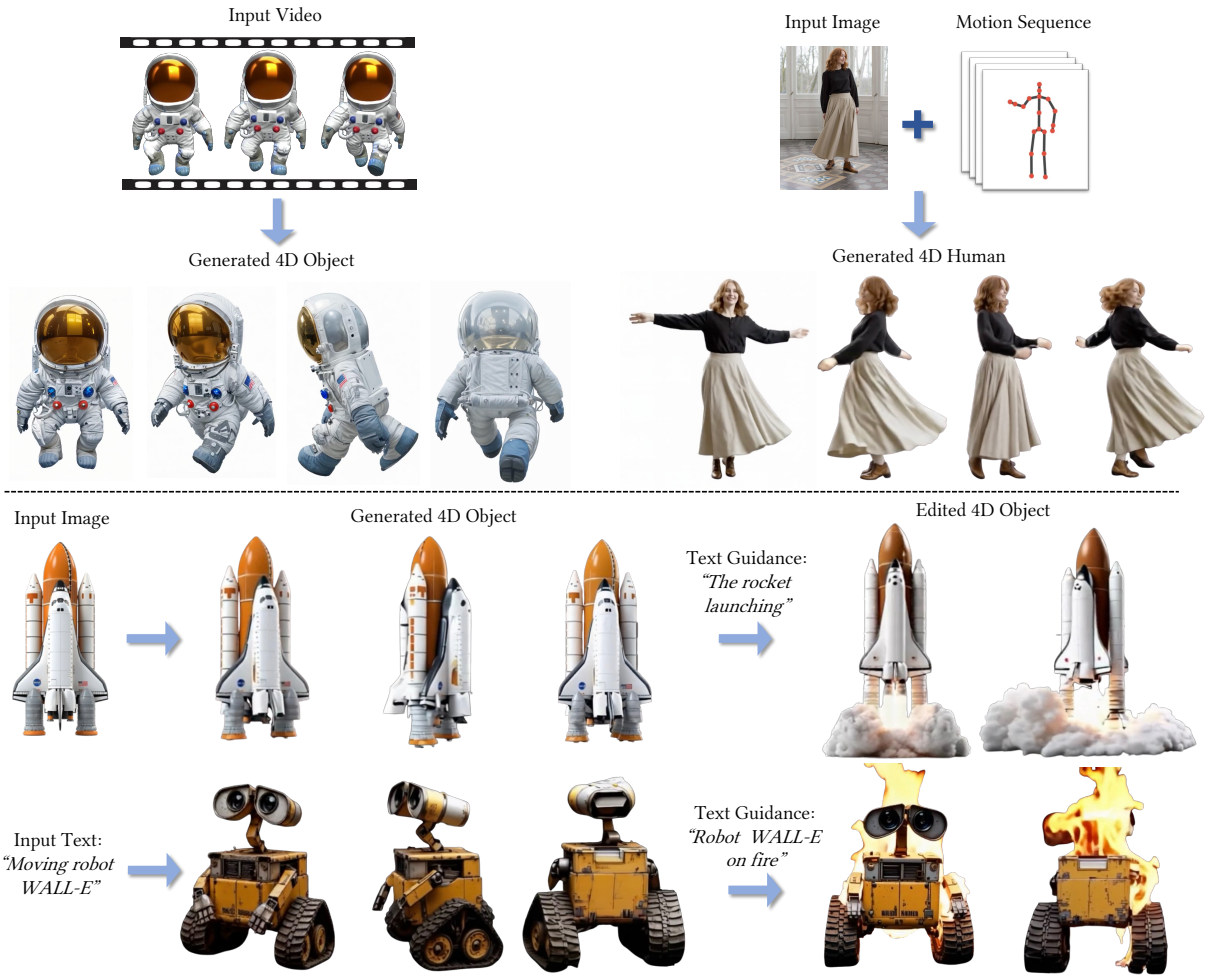


Fig. 1. **Splat4D.** Our method empowers a wide array of 4D content generation capabilities with high fidelity. Top-left: Generating the 4D representation from a monocular video; Top-right: 4D human generation guided by an image and motion sequence; Bottom-left: Generation of dynamic 4D objects from image or text inputs; Bottom-right: Text-guided 4D content editing, enabling detailed scene customization.

*Corresponding author is Kai Han (kaihanx@hku.hk).

Authors' Contact Information: Minghao Yin, The University of Hong Kong, Hong Kong; Yukang Cao, Nanyang Technological University, Singapore; Songyou Peng, Google DeepMind, USA; Kai Han, The University of Hong Kong, Hong Kong.


*SIGGRAPH Conference Papers '25, Vancouver, BC, Canada*

ACM ISBN 979-8-4007-1540-2/2025/08
https://doi.org/10.1145/3721238.3730752

Generating high-quality 4D content from monocular videos—for applications such as digital humans and AR/VR—poses challenges in ensuring temporal and spatial consistency, preserving intricate details, and incorporating user guidance effectively. To overcome these challenges, we introduce Splat4D, a novel framework enabling high-fidelity 4D content generation from a monocular video. Splat4D achieves superior performance while maintaining faithful spatial-temporal coherence, by leveraging multi-view rendering, inconsistency identification, a video diffusion model, and an asymmetric U-Net for refinement. Through extensive evaluations on public benchmarks, Splat4D consistently demonstrates state-of-the-art performance across various metrics, underscoring the efficacy of our approach. Additionally, the versatility of Splat4D is validated in various applications such as text/image conditioned 4D generation, 4D human generation, and text-guided content editing, producing coherent outcomes following user instructions. Project page: https://visual-ai.github.io/splat4d

CCS Concepts: • **Computing methodologies** → *Animation*; *Reconstruction*.

Additional Key Words and Phrases: 4D Generation, Gaussian Splatting

**ACM Reference Format:**
Minghao Yin, Yukang Cao, Songyou Peng, and Kai Han. 2025. Splat4D: Diffusion-Enhanced 4D Gaussian Splatting for Temporally and Spatially Consistent Content Creation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25), August 10–14, 2025, Vancouver, BC, Canada.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3721238.3730752

## 1 Introduction

The generation of 4D content—encompassing dynamic 3D objects—is integral to applications in digital humans, gaming, media, and AR/VR, where realistic motion and spatial-temporal consistency are essential. Unlike static 3D object generation [Chen et al. 2023; Lin et al. 2023; Melas-Kyriazi et al. 2023; Qian et al. 2023; Raj et al. 2023; Tang et al. 2024b, 2023; Voleti et al. 2024; Wang et al. 2023], creating 4D content must capture an object's evolving appearance and motion within 3D space, which significantly increases the complexity. This challenge is especially pronounced when generating 4D content from a single monocular video, as it demands simultaneous inference of appearance and motion for unseen camera viewpoints. Moreover, the problem is inherently ill-posed, as multiple valid 4D interpretations can emerge from the same input. Consequently, representing 3D shape, texture, and motion in this high-dimensional space requires a substantial number of parameters, emphasizing the need for efficient modeling and computational strategies to address these demands.

Recent works [Bahmani et al. 2024; Cao et al. 2024b; Ren et al. 2023; Singer et al. 2023b; Wang et al. 2024; Zhao et al. 2023; Zheng et al. 2024] have explored 4D content generation through score-distillation sampling (SDS) [Poole et al. 2023] loss with pre-trained diffusion models, producing dynamic scenes but often suffering from slow generation speeds and spatial-temporal inconsistencies. To address these limitations, follow-up approaches [Jiang et al. 2023; Yin et al. 2023; Zeng et al. 2024] leverage 3D-aware diffusion models [Shi et al. 2023] to improve spatial consistency. Recently, with the development of video diffusion models [Blattmann et al. 2023a,b; He et al. 2022], different techniques have been explored for fine-tuning these models to enable the generation of multi-view video sequences from single-view inputs. Leveraging this enhanced multi-view priors, methods such as 4Diffusion [Zhang et al. 2024], SV4D [Xie et al. 2024], and Diffusion4D [Liang et al. 2024b] have been proposed to advance 4D generation and reconstruction more efficiently. Despite these advancements, significant challenges persist, including ensuring temporal and spatial consistency, accurately modeling complex human characteristics (e.g., loose clothing), and effectively integrating diverse input modalities such as images, text, and motion data.

To address these persistent challenges, we propose a novel framework for *generalizable 4D content creation*, called Splat4D, which allows for high-quality 4D generation from a monocular video input, rendering versatile applications. First, we transform a monocular video into high-quality multi-view image sequences with a multi-view diffusion model and an image enhancer. A pretrained large-scale feed-forward 3DGS model is then employed to obtain spatial and depth features across views, which are then further processed by Splatter Image to produce a comprehensive yet coarse Gaussian field. Next, we introduce a method for refining spatial and temporal coherence of the 4D Gaussian field representation through multi-view rendering, inconsistency identification, and a video diffusion model, resulting in a high-quality 4D representation with improved visual quality and stability. Finally, to boost the realism of the generation, an asymmetric U-Net model is trained as the generalizable 3D Gaussian field predictor to produce accurate and detailed 3D Gaussians, improving overall quality. Our Splat4D model is the first to incorporate an image enhancer for high-quality 4D generation, leading to substantial performance gains. However, directly applying the enhancer can cause multi-view inconsistencies and misalignments. To address this, we introduce uncertainty masking and asymmetric U-Net training to identify unreliable regions and adaptively refine the reconstruction. The video diffusion model then inpaints the masked areas, ensuring spatial-temporal consistency. These components are carefully integrated to complement each other, achieving results unattainable by any single module alone. By seamlessly integrating these components, our Splat4D framework can effectively generate high-quality spatial-temporal consistent 4D content at speed. Meanwhile, it can be applied for various of applications, such as text/image conditioned 4D generation, 4D human generation, and text-guided content editing (see Fig. 1).

We thoroughly evaluate our framework on public benchmark, achieving the state-of-the-art results across the board on all metrics, validating the superior performance of our method. Moreover, we also showcase the results of applying our method for the applications of text/image conditioned 4D generation, 4D human generation, and text-guided content editing, demonstrating faithful and coherent results following the provided guidance.

## 2 Related Work

**3D Generation.** For 3D content generation, early works such as DreamFusion [Poole et al. 2023] pioneer the use of Score Distillation Sampling (SDS) loss to distill priors from 2D diffusion models, optimizing 3D content from textual or image input. Subsequent efforts [Cao et al. 2023, 2024a; Chen et al. 2024; Han et al. 2023; Li et al. 2024; Liang et al. 2024a; Pan et al. 2024; Sargent et al. 2024; Sun et al. 2023; Tang et al. 2024b; Wang et al. 2023; Weng et al.

2023; Yi et al. 2024; Zhou et al. 2024] address challenges like multi-view Janus artifacts, slow generation speed, and oversaturation by fine-tuning diffusion models for viewpoint control or directly generating multi-view images within a single diffusion pass. Methods like Zero123 [Liu et al. 2023] and SyncDreamer [Liu et al. 2024] refine 2D diffusion models for multi-view consistency, while others, including Magic3D [Lin et al. 2023] and Direct2.5 [Lu et al. 2024], adopt alternative 3D representations such as Instant-NGP [Müller et al. 2022], DMTet [Müller et al. 2022], or explicit mesh-based approaches [Lu et al. 2024] to improve runtime and fidelity.

DreamGaussian [Tang et al. 2024b] introduces a point-based representation, utilizing 3D Gaussians for faster generation and superior quality compared to traditional Neural Radiance Fields (NeRF) [Mildenhall et al. 2021]. The feed-forward method LGM (Large Multi-View Gaussian Model) [Tang et al. 2024a] efficiently represents scenes with multi-view Gaussian features and uses an asymmetric U-Net to process multi-view images. In our work, we draw inspiration from the ideas introduced in DreamGaussian [Tang et al. 2024b], which uses 3D Gaussians for faster, high-quality generation, and LGM [Tang et al. 2024a], which employs multi-view Gaussian features and asymmetric U-Nets to process multi-view images from single-view inputs.

**Video Diffusion Model.** Recent video diffusion models have achieved impressive results in creating realistic motions and geometrically consistent sequences [Blattmann et al. 2023a,b; Guo et al. 2023; He et al. 2022; Ho et al. 2022; Singer et al. 2023a; Voleti et al. 2022]. Their strong generalization abilities stem from training on extensive image and video datasets, which are more accessible than large-scale 3D or 4D datasets. These models are increasingly utilized as foundational tools for tasks like multi-view synthesis and 3D content generation. For instance, frameworks like SV3D [Voleti et al. 2024] adapt latent video diffusion models to produce consistent multi-view imagery, while approaches like AnimateDiff [Guo et al. 2023] enhance text-to-image models by incorporating motion modules to capture temporal dynamics. Similarly, SV4D [Xie et al. 2024] employs video diffusion models to achieve both video generation and novel view synthesis. Leveraging these advancements, our approach extends pre-trained video generation models by introducing view attention mechanisms, aligning outputs for improved coherence in multi-view and 4D applications.

**4D Generation.** DreamGaussian4D [Ren et al. 2023] leverages a three-stage framework to generate high-quality 4D animations. It uses a modified version of Gaussian splatting combined with image-to-video diffusion for high-fidelity 3D static models that are deformed over time using a learned deformation field. DYST (Dynamic Scene Transformer) [Seitzer et al. 2023] innovates further by decomposing monocular videos into distinct latent representations of scene content, per-view dynamics, and camera pose through co-training on real and synthetic data. GaussianFlow [Gao et al. 2024] introduces Gaussian fields paired with optical flow constraints, further enhancing consistency in generated motion by aligning temporal transitions. AvatarGO [Cao et al. 2024b] proposes a correspondence-aware motion field that enables harmonious generation of 4D human–object interactions from text.

Building on the foundation of 3D-aware diffusion models [Shi et al. 2023], recent methods such as Stag4D [Zeng et al. 2024] and 4DGen [Yin et al. 2023] focus on enhancing spatial consistency in 4D generation. Meanwhile, Consistent4D [Jiang et al. 2023] employs a video interpolation model to improve both temporal and spatial coherence. More recently, video diffusion models have been adopted for further 4D content enhancement. For instance, 4Diffusion [Zhang et al. 2024] introduces a multi-view video diffusion model coupled with a 4D-aware SDS loss to optimize dynamic NeRF representations. Similarly, Diffusion4D [Liang et al. 2024b] leverages a 4D-aware video diffusion framework combined with explicit 4D construction to synthesize 4D assets. For feed-forward generation, SV4D [Xie et al. 2024] advances this line of work by utilizing a multi-view video synthesis model for efficient 4D optimization. Additionally, L4GM [Ren et al. 2024] proposes a 4D interpolation model enabling fast feed-forward 4D generation from single-view video inputs.

However, current 4D generation methods often struggle to produce high-quality content that maintains both spatial and temporal consistency. Issues such as blurry textures, geometric distortions, and temporal flickering are still prevalent. To address these limitations, we introduce Splat4D, a novel approach that can generate high-quality, spatial-temporal consistent 4D content.

## 3 Method

Given a single-view image $I$ or a text prompt $y$ as input, Splat4D facilitates the generation of a 4D dynamic scene. Specifically, our method captures both the spatial structure and the temporal evolution of the scene by decomposing it into multiple 3D Gaussian distributions. In the subsequent sections, we first illustrate the preliminaries that underpin our method in Section 3.1. We then delve into the core techniques of our method, including (1) coarse 4D Gaussian generation that leverages pre-trained diffusion priors in Section 3.2, (2) temporal and spatial refinement for improving consistency in Section 3.3, and (3) generalizable 3D Gaussian field predictor learning in Section 3.4. An overview of our pipeline is shown in Fig. 2.

### 3.1 Preliminaries

*3.1.1 3D Gaussian Splatting.* Different from NeRF [Mildenhall et al. 2021], which relies on neural networks for novel view synthesis, 3D Gaussian Splatting (3DGS) [Kerbl et al. 2023] takes a different approach by directly optimizing the 3D position $\mathbf{x}$ and attributes of 3D Gaussians, such as opacity $\alpha$, anisotropic covariance, and spherical harmonic (SH) coefficients $\mathcal{SH}$ [Ramamoorthi and Hanrahan 2001]. Each 3D Gaussian $G(\mathbf{x})$ is characterized by a 3D covariance matrix $\Sigma$ centered at its mean position $\mu$.

$$G(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}. \tag{1}$$

For 3DGS, a tile-based rasterizer is utilized by dividing the screen into tiles, such as $16\times16$ pixels. For each tile it intersects, a Gaussian is instantiated and assigned a key encoding its depth in view space and the corresponding tile ID. Depth sorting is then applied to the Gaussians, enabling the rasterizer to efficiently resolve occlusions and overlapping structures. The final RGB color $\mathbf{C}$ is computed using

a point-based $\alpha$-blending approach, which samples points along the ray at regular intervals.

*3.1.2 Large Multi-View Gaussian Model (LGM).* The LGM framework (e.g. [Tang et al. 2024a]) transforms a single input image into a 3D Gaussian representation of the object through a systematic process. First, it employs MVDream [Shi et al. 2023], a multi-view diffusion model, to generate multi-view consistent images from the input. MVDream synthesizes images from four viewpoints, ensuring they align geometrically and maintain visual consistency. These multi-view images are then processed by an asymmetric U-Net, the encoder extracts multi-scale spatial features from the images, while the decoder focuses on reconstructing these features into dense splatter images [Szymanowicz et al. 2024]. Splatter images encode the parameters for 3D Gaussians for each pixel, such as position, size, color, and opacity. These splatter images are then transformed into 3D Gaussian representations by backprojecting the pixel-wise Gaussian parameters into 3D space using the camera parameters. We leverage the feedforward LGM for efficient but coarse scene representation generation.

## 3.2 Coarse 4D Gaussian Generation

Existing 4D generative methods [Bahmani et al. 2024; Ren et al. 2023; Singer et al. 2023b; Wang et al. 2024; Zhao et al. 2023; Zheng et al. 2024] typically employ pre-trained diffusion models [He et al. 2022] and score distillation sampling (SDS) [Poole et al. 2023] to subsequently generate and animate 3D scenes from either text or image guidance. However, we observe two main limitations in such techniques: (1) it struggles to produce large, dynamic motions effectively, and (2) it requires significant training time to generate a single outcome. Drawing inspiration from the success of SV4D [Xie et al. 2024], our approach aims to overcome these limitations by first generating single-view videos and then distilling these videos into the 4D space for further refinement and animation.

*3.2.1 Multi-view Video Generation.* We utilize the video diffusion model [Blattmann et al. 2023a] to generate the image sequence. However, relying solely on a single-view video does not provide enough information for robust 4D modeling. This limitation stems from issues like depth ambiguity and the lack of side and back view information. To address this, we enhance the single-view video by using MV-Adapter [Huang et al. 2024] to generate additional views, including the front, back, and sides, thereby enriching the model with more comprehensive rotational perspectives:

$$\texttt{MV-Adapter}(I_t) \rightarrow \{I_t, I_t^{\text{left}}, I_t^{\text{right}}, I_t^{\text{back}}\}, \tag{2}$$

See Supplementary Material for evaluation for the choice of MV-Adapter over SV4D.

*3.2.2 Multi-view Image Enhancer.* Although MV-Adapter can robustly provide multi-view perspectives, their generated videos often lack fine-grained details and the high resolution required for realistic 4D content (see the figure in the supplementary material). This issue is expected as the input samples would always fall outside the training distribution of MV-Adapter. To address this, we propose to apply an image enhancer model [Wang et al. 2018] `IE` to refine textures, edges, and details for each frame and each view.

*3.2.3 4D Gaussian Reconstruction.* After generating a high-quality, multi-view image sequence, we proceed to construct a 4D Gaussian field. Following LGM [Tang et al. 2024a], we first input the multi-view image sequence into U-Net to encode key spatial and depth features across the views. The U-Net architecture is well-suited for this task because it can capture detailed structures at multiple resolutions through its encoder-decoder structure. The encoder captures feature maps at different scales, identifying essential textures and depth cues, while the decoder reconstructs these features into a cohesive representation.

Once the U-Net has processed the multi-view sequence, we apply the Splatter Image [Szymanowicz et al. 2024] method to project these learned features into a continuous 4D Gaussian field. Specifically, Splatter Image maps each pixel from the feature maps into a series of localized Gaussian distributions in 3D space, with each Gaussian representing a small spatial region from the scene. To form the final temporally consistent Gaussian sequences, we design our network to separately reconstruct a 3D Gaussian field for each frame (time step). Specifically, we construct a stacked representation of multiple 3D Gaussians, represented as $\mathcal{G}(\mathcal{S}, t) = [\mathcal{X}_t, s_t, r_t, \sigma_t, \zeta_t]$, with position, scale, rotation, opacity and Spherical Harmonics (SH) at time $t$. This Gaussian field serves as a foundational structure, providing a spatially continuous and temporally stable representation that can be rendered from any angle.

## 3.3 Spatial-Temporal Consistency Refinement

Although multi-view video generation and image enhancement techniques can provide detailed 3D information necessary for constructing a 4D Gaussian scene, the resulting reconstruction still suffers from issues with temporal and spatial consistency. This happens because the MV-Adapter has difficulty maintaining consistent multi-view images. Additionally, since the MV-Adapter processes each frame independently, it further contributes to these inconsistencies in the model. To tackle this problem, we introduce a multi-step approach that involves two key techniques: *inconsistency masking* and *uncertainty-guided refinement*.

*3.3.1 Inconsistency Masking.* We start by rendering a sequence of multi-view images $\{\mathcal{I}_t, \mathcal{I}_t^{\text{left}}, \mathcal{I}_t^{\text{right}}, \mathcal{I}_t^{\text{back}} | t \in [1, T]\}$ from the 4D Gaussian field $\mathcal{G}(\mathcal{S}, t)$, where $\mathcal{I}$ represents the rendered images. For each time step $t$, we then generate uncertainty maps [Kulhanek et al. 2024] to detect regions with inconsistencies. We extract DINOv2 [Oquab et al. 2024] features from the rendered images and predict the pixel-wise uncertainty $\sigma$ using an uncertainty prediction network [Kulhanek et al. 2024]. These uncertainty maps highlight areas that show significant variation or deviation between frames, which are often caused by issues like motion artifacts, occlusions, or perspective differences. By identifying these inconsistent areas, we create a mask that helps us focus on correcting the problematic regions while keeping the stable areas intact. The uncertainty mask is defined as $M = \mathbf{1}\left(\frac{1}{2\sigma^2} > 1\right)$, where $\mathbf{1}$ is the indicator function.

*3.3.2 Uncertainty-guided Refinement.* Inspired by [Yu et al. 2024], we address the inconsistencies highlighted by the uncertainty map by applying a video denoising diffusion model [Xing et al. 2024] to the rendered sequence. This model leverages the masked areas

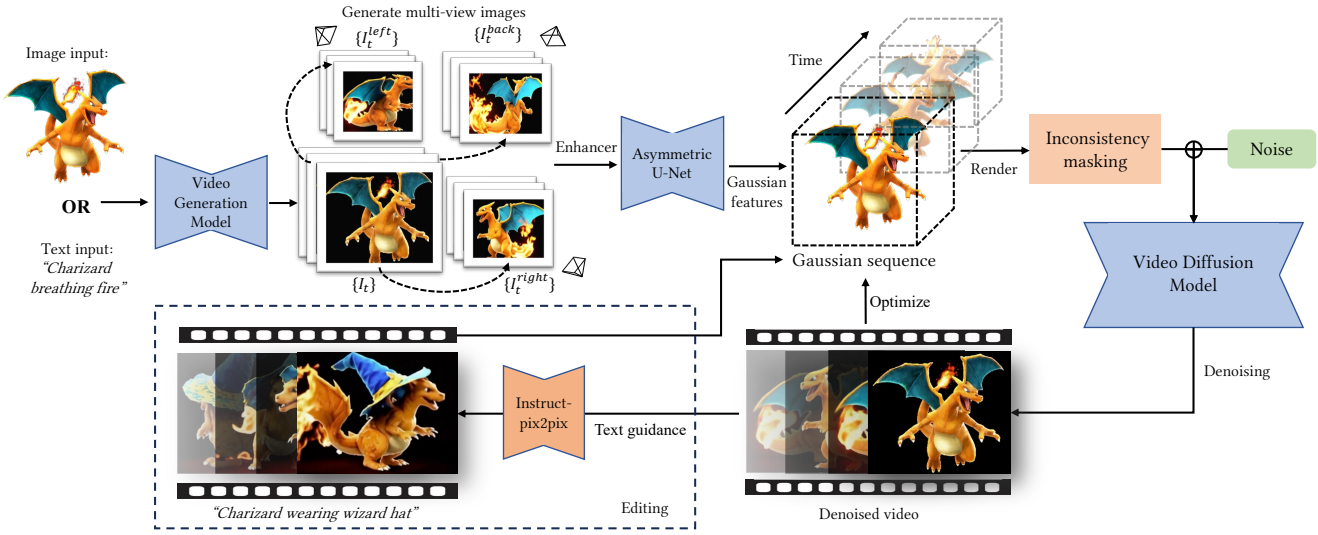


Fig. 2. **Overview of Splat4D**. Our method for 4D content generation begins with processing input data (text, image, or monocular video) to produce high-quality multi-view image sequences. These sequences are used to initialize a 4D Gaussian representation via an asymmetry U-Net and image splattering. Refinement steps include leveraging uncertainty masking and video denoising diffusion to ensure high fidelity and spatial-temporal consistency, culminating in versatile 4D content creation. The pipeline supports optional text-guided content editing, enabling dynamic modifications of the 4D output for enhanced flexibility and creative control.

identified earlier and restores the temporal and spatial consistency by "filling in" these regions with content that aligns seamlessly with the surrounding pixels. The diffusion model operates iteratively, refining each frame while considering the neighboring frames to ensure smooth transitions and maintain consistent visual quality. This step is crucial for preserving the flow of the sequence, reducing issues like jitter or flicker that can disrupt the viewer's experience. Once the sequence is refined and consistent, the updated frames are used to improve the 4D Gaussian field. This creates a feedback loop, aligning the 4D representation with the enhanced image sequence, which boosts the overall quality and stability of the 4D scene. Note that we condition the video diffusion model on the first and last frames of the input sequence to address the hallucination problem.

### 3.4 Generalizable 3D Gaussian Field Predictor Learning

Aside from the inconsistency issues we've already improved, we also find that the quality of the 4D Gaussian fields doesn't always match the improvements made by the image enhancer. This is expected as there is a notable domain gap between the pre-trained distribution of the image enhancer model [Wang et al. 2018], which is trained on DIV2K dataset [Agustsson and Timofte 2017], and LGM [Tang et al. 2024a], which is trained on Objaverse [Deitke et al. 2023]. To address this issue, we propose to fine-tune the U-Net model derived from LGM with the pre-processed Objaverse dataset. Specifically, we first follow LGM [Tang et al. 2024a] to filter low-quality 3D models. In each training step, we randomly choose an input image with an elevation angle between -30 and 30 degrees. The MV-Adapter [Huang et al. 2024] is then used to generate four orthogonal views, including the original image. These views are processed through the image enhancer and consistency refinement steps, and are subsequently passed into the U-Net model to produce the 4D Gaussian field. Finally, we render images from the Gaussian field based on the angles of the four orthogonal views for supervision. This training process allows the fine-tuned U-Net to reduce the domain gap between the pre-trained U-Net from LGM and the image enhancer model, resulting in improved quality of the 4D Gaussian fields.

## 4 Experiments

For the experiments, we conduct both qualitative and quantitative comparisons on 4D generation, perform ablation studies, and explore various applications of our method.

### 4.1 Implementation Details

For the evaluation of video-to-4D generation, we utilize the video dataset provided by Consistent4D [Jiang et al. 2023]. We employ the Segment Anything Model (SAM) [Kirillov et al. 2023] to preprocess the input image sequences to extract the foreground objects. To evaluate image-to-4D generation, we curate a dataset by collecting images from the internet. These images are converted to RGBA format and resized to a resolution of 512×512 to ensure compatibility with our pipeline. For fine-tuning, we utilize the 80K 3D object subset[Tang et al. 2024a] of the Objaverse dataset [Deitke et al. 2023] after filtering out low-quality models. Each 3D model is rendered into RGB images from 100 camera views at a resolution of $512 \times 512$.

The training process is being conducted using the asymmetric U-Net model on 4 NVIDIA V100 GPUs, with each GPU processing a batch size of 4 under bfloat16 precision. For each batch, a single camera view is being randomly sampled, while 4 orthogonal views are being generated using the MV-adapter [Huang et al. 2024] based on the input view. The asymmetric U-Net model is generating the

Table 1. ***Video-to-4D*** **quantitative Comparison on Consistent4D Dataset [Jiang et al. 2023].**

| Model | LPIPS↓ | CLIP-S↑ | FVD-F↓ | FVD-V↓ |
|---|---|---|---|---|
| Consistent4D | 0.134 | 0.87 | 1133.93 | 735.79 |
| STAG4D | 0.126 | 0.91 | 992.21 | 685.23 |
| SV4D | 0.118 | 0.92 | 732.40 | 503.51 |
| 4Diffusion | 0.13 | 0.94 | 489.2 | 405.5 |
| **Ours** | **0.090** | **0.97** | **390.85** | **282.79** |

Table 2. ***Video-to-4D*** **Quantitative Comparison on ObjaverseDy Test Set [Deitke et al. 2023; Xie et al. 2024].**

| Model | LPIPS↓ | CLIP-S↑ | FVD-F↓ | FVD-V↓ |
|---|---|---|---|---|
| Consistent4D | 0.165 | 0.896 | 880.54 | 488.38 |
| STAG4D | 0.158 | 0.860 | 929.10 | 453.62 |
| SV4D | 0.131 | 0.905 | 659.66 | 368.53 |
| **Ours** | **0.112** | **0.939** | **383.71** | **267.94** |

Table 3. **Quantitative Comparison on *Image-to-4D* Generation.**

| Model | LPIPS↓ | CLIP-S↑ | PSNR↓ | FVD↓ |
|---|---|---|---|---|
| 4DGen | 0.28 | 0.84 | 14.4 | 736.6 |
| STAG4D | 0.24 | 0.86 | 15.2 | 675.4 |
| Diffusion4D | 0.18 | 0.89 | 16.8 | 490.2 |
| **Ours** | **0.12** | **0.94** | **19.2** | **395.0** |

Table 4. **Ablation Study.** The experiments are conducted on Consistent4D dataset [Jiang et al. 2023].

| Model | LPIPS↓ | CLIP-S↑ | FVD-F↓ | FVD-V↓ |
|---|---|---|---|---|
| w/o mask | 0.114 | 0.93 | 507.15 | 413.79 |
| w/o train | 0.107 | 0.96 | 445.33 | 364.81 |
| **Ours** | **0.090** | **0.98** | **390.85** | **282.79** |

3D Gaussian field, which is then being rendered into images for the orthogonal views. Original Objaverse 3D object rendered images are being used as supervision signals. The rendered 3D Gaussians are being compared to the original at a resolution of 512×512 using the mean squared error (MSE) loss. To optimize memory usage, images are being resized to 256×256 for LPIPS loss calculation. The AdamW optimizer is being employed with a learning rate of $4 \times 10^{-4}$, a weight decay of 0.05, and momentum parameters of 0.9. The learning rate is following a cosine annealing schedule to gradually decay to zero during training. Gradients are being clipped to a maximum norm of 1.0 to enhance stability. Additionally, grid distortion and camera jitter are being applied with a probability of 50% to improve generalization.

## 4.2 Main Comparison

We compare our model with state-of-the-art baselines including STAG4D [Zeng et al. 2024], SV4D [Xie et al. 2024], Consistent4D [Jiang et al. 2023], 4Diffusion [Zhang et al. 2024] and Diffusion4d [Liang et al. 2024b]. As shown in Fig. 3, both STAG4D and SV4D suffer from producing satisfactory results when rendering novel views. This distinction is particularly pronounced in scenarios involving complex human structures, such as detailed facial features and loose clothing (first row). These results underscore the superior capability of our method in performing robust video-to-4D reconstruction.

For *video-to-4D* quantitative evaluation in Table 1, we assess the quality of each generated image by comparing it with its corresponding ground truth using metrics such as Learned Perceptual Similarity (LPIPS) and CLIP-Score (CLIP-S). These metrics help evaluate the visual fidelity and semantic alignment of the generated images. To measure temporal and spatial coherence in the generated video, we report the Fréchet Video Distance (FVD), a widely used video-level metric in video generation tasks. Our method achieves the best performance compared to all baselines across all evaluation metrics on Consistent4D dataset [Jiang et al. 2023] and ObjaverseDy test set [Deitke et al. 2023; Xie et al. 2024] as shown in Table 1, indicating that our method is able to generate temporally and spatially coherent 4D content. For *image-to-4D* evaluation in Table 3, we compare the generation results from measurements as in [Liang et al. 2024b]. Our method still outperforms all baseline methods.

## 4.3 Ablation Study

We conduct an ablation study on the Consistent4D dataset [Jiang et al. 2023] and ObjaverseDy test set [Deitke et al. 2023; Xie et al. 2024], to assess the impact of key components in our method. Specifically, we examine three variants: no uncertainty masking, 2) no U-Net training, and 3) our full model. The evaluation metrics include LPIPS, CLIP-S, FVD-F, and FVD-V, which respectively measure perceptual similarity, alignment with textual semantics, and spatial-temporal consistency.

As shown in Table 4, omitting either the uncertainty masking or U-Net training *significantly degrades* LPIPS, CLIP-S and FVD metrics, demonstrating the importance of uncertainty map both components in handling spatial-temporal inconsistencies and preserving high-fidelity. In Supplementary Material, we present additional comparisons of model design choices, including the use of feedback loop, the incorporation of the image enhancer, and video diffusion model conditioning on first/last frames as discussed in Section 3.3.

## 4.4 Splat4D for Different Applications

Besides taking a single-view image or a text prompt as input to obtain 4D dynamic scenes, our method can also be applied to different tasks, including 4D human motion transfer, and text-guided 4D content editing.

*4.4.1 Text/Image Conditioned 4D Generation.* To demonstrate our method's capability for *text-to-4D* and *image-to-4D* generation, we show our generation results in Fig. 4. For *text-to-4D generation*, we first employ a text-to-image diffusion model to convert the input textual prompt into a high-quality image and then combine that image with the original text in a stable video diffusion model to produce a coherent short video. In *image-to-4D generation*, the pipeline

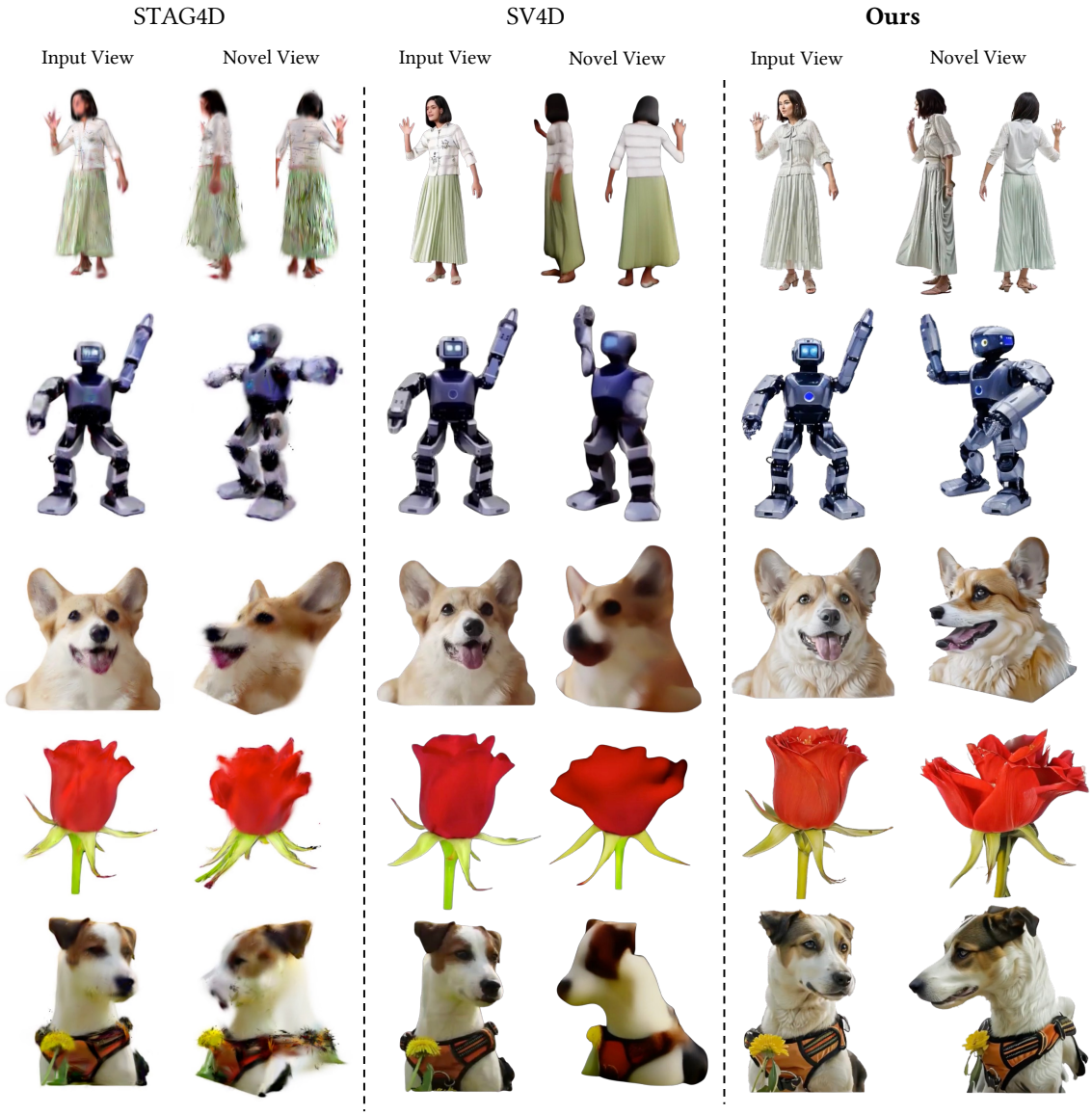


Fig. 3. **Comparison on *video-to-4D* generation**. The rendered image of the input view from the 4D object is on the left column, and the rendered images of the novel view are illustrated on the two right columns.

simply begins with an input image, bypassing the text-to-image step. From the resulting videos, we use a multi-view diffusion model to generate four orthogonal view sequences, which are then fed into our reconstruction pipeline to construct a 4D Gaussian field.

As shown in Fig. 4, the top two rows illustrate *image-to-4D* results, while the bottom two rows depict *text-to-4D* outputs. The generated 4D reconstructions demonstrate the effectiveness of our pipeline in maintaining structural coherence and high visual fidelity across multiple perspectives. For the *image-to-4D* examples, we observe precise alignment and consistent detail retention in novel views derived from the input image. As for the *text-to-4D* results, the generated scenes accurately align with the semantic content of the input text, producing dynamic and visually plausible outputs.

*4.4.2 4D Human Generation.* Given a source video with the desired motion to be transferred and an image of the human subject to be animated, we first use a pose extraction model [Kocabas et al. 2020] to detect key body landmarks, skeletal poses, and motion trajectories. Next, we apply Champ [Zhu et al. 2024], a 2D motion transfer model, to animate the input image, making it move according to the extracted motion. Our method then uses the resulting animated image sequence $\{I_t | t \in [1, T]\}$, where $T$ is the total number of frames, to generate the corresponding 4D Gaussian scenes. Fig. 5

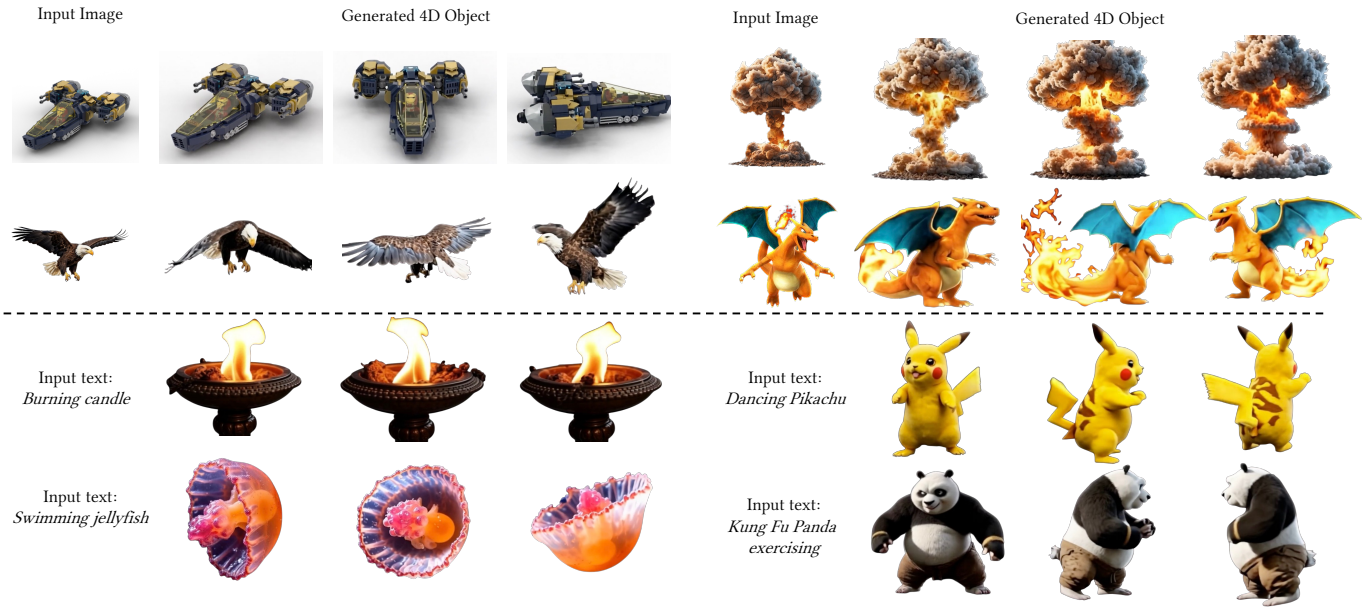


Fig. 4. **4D Content Creation with Text/Image as Input**. The first two rows are results with image inputs, and bottom two rows are results with text inputs.

showcases the results of our *4D human generation* pipeline, which combines a single input image with a motion sequence to produce high-fidelity, dynamic human representations. First, the input image and motion sequence are processed through a 2D motion transfer model to create a video of the subject performing the specified action (see details in Supplementary Material). Next, we follow our pipeline and apply a multi-view diffusion model and construct 4D Gaussians of the human.


Fig. 5. **4D Human Generation with an Input Image as Guidance**. The first row shows the input image, while the subsequent rows depict rendered novel views under various poses.


Fig. 6. **4D Content Editing with Text Guidance**. The first column showcases the original input (text or image), while the subsequent three columns present the edited outputs. Each edited 4D object is displayed beneath the corresponding text.

Fig. 5 illustrates the effectiveness of our approach. These results highlight our ability to preserve intricate human details, including complex structures like facial details and loose clothing, across varying views and motions.

*4.4.3 Text-guided Editing.* For this application, we use the Instruct-Pix2Pix [Brooks et al. 2023] network to modify the output video

generated by the video diffusion model, guided by a text prompt (see Fig. 2). For instance, starting with a video of a house, the pix2pix network is applied to transform the video based on a prompt like "house on fire". Specifically, the pix2pix network performs image-to-image translation, adjusting each frame of the video to match the specified scene changes. Once the transformation is complete, the modified video sequence is used to refine the corresponding 4D Gaussian sequence, resulting in a final 4D content that accurately reflects the updated dynamics of the "house on fire" scenario. In Fig. 6, we showcase our method's *text-guided 4D editing* capability that transforms 4D Gaussian representations based on user-specified textual or visual prompts. Starting from our 4D Gaussian field, we employ a pix2pix network to edit the rendered video according to the guidance text, producing an updated video sequence. This sequence is further optimized using the 4D Gaussian representation, ensuring coherence and alignment with the guidance.

The results are illustrated in Fig. 6. These examples demonstrate the effectiveness of our approach in introducing realistic and coherent transformations, such as attribute changes or new dynamic effects, while maintaining high fidelity to the original 4D content structure.

## 5 Conclusions

In this paper, we introduce a novel framework for high-quality 4D content generation, which addresses key challenges in dynamic scene creation by leveraging a 4D Gaussian splatting representation. Our method demonstrates strong generalization capabilities, enabling the generation of temporally stable and high-fidelity 4D content from monocular videos, images, and text prompts. Through careful integration of a multi-view video diffusion model and an asymmetric U-Net, we improve both spatial and temporal consistency, enhancing the visual coherence of the generated scenes. Our ablation studies validate the importance of components such as uncertainty map masking and asymmetry U-Net training for improving the quality of 4D content generation. The proposed framework is versatile and can be applied to a variety of scenarios, including text/image conditioned 4D generation, 4D human generation, and text-guided content editing. We believe that our approach marks a significant advancement in 4D scene generation, offering a robust solution that balances computational efficiency with high-quality results for real-world applications in digital humans, gaming, AR/VR, and media production.

## Acknowledgments

The work is supported by the Hong Kong Research Grants Council - General Research Fund (Grant No.: 17211024).

## References

Eirikur Agustsson and Radu Timofte. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 126–135.

Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 2024. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7996–8006.

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

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22563–22575.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.

Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. 2023. Guide3d: Create 3d avatars from text and image guidance. *arXiv preprint arXiv:2308.09705* (2023).

Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. 2024a. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 958–968.

Yukang Cao, Liang Pan, Kai Han, Kwan-Yee K Wong, and Ziwei Liu. 2024b. Avatargo: Zero-shot 4d human-object interaction generation and animation. *arXiv preprint arXiv:2410.07164* (2024).

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22246–22256.

Zilong Chen, Feng Wang, Yikai Wang, and Huaping Liu. 2024. Text-to-3d using gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21401–21412.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13142–13153.

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

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2023. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *The Twelfth International Conference on Learning Representations*.

Xiao Han, Yukang Cao, Kai Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, Tao Xiang, and Kwan-Yee K Wong. 2023. Headsculpt: Crafting 3d head avatars with text. *Advances in neural information processing systems* 36 (2023), 4915–4936.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. 2022. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221* (2022).

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *Advances in Neural Information Processing Systems* 35 (2022), 8633–8646.

Zehuan Huang, Yuan-Chen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. 2024. MV-Adapter: Multi-view Consistent Image Generation Made Easy. *arXiv preprint arXiv:2412.03632* (2024).

Yanqin Jiang, Li Zhang, Jin Gao, Weiming Hu, and Yao Yao. 2023. Consistent4D: Consistent 360° Dynamic Object Generation from Monocular Video. In *The Twelfth International Conference on Learning Representations*.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. 2020. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5253–5263.

Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. 2024. WildGaussians: 3D Gaussian Splatting in the Wild. In *Advances in Neural Information Processing Systems*, Vol. 37.

Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. 2024. SweetDreamer: Aligning Geometric Priors in 2D diffusion for Consistent Text-to-3D. In *The Twelfth International Conference on Learning Representations*.

Hanwen Liang, Yuyang Yin, Dejia Xu, hanxue liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. 2024b. Diffusion4D: Fast Spatial-temporal Consistent 4D generation via Video Diffusion Models. In *Advances in Neural Information Processing Systems*, Vol. 37. 110854–110875.

Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. 2024a. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6517–6526.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 300–309.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9298–9309.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. 2024. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. In *The Twelfth International Conference on Learning Representations*.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 851–866.

Yuanxun Lu, Jingyang Zhang, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, Xun Cao, and Yao Yao. 2024. Direct2. 5: Diverse text-to-3d generation via multi-view 2.5 d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8744–8753.

Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. 2023. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8446–8455.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)* 41, 4 (2022), 1–15.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research Journal* (2024), 1–31.

Zijie Pan, Jiachen Lu, Xiatian Zhu, and Li Zhang. 2024. Enhancing High-Resolution 3D Generation through Pixel-wise Gradient Clipping. In *The Twelfth International Conference on Learning Representations*.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *The Eleventh International Conference on Learning Representations*.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10318–10327.

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

Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. 2023. Dreambooth3d: Subject-driven text-to-3d generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2349–2359.

Ravi Ramamoorthi and Pat Hanrahan. 2001. An efficient representation for irradiance environment maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. 497–500.

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

Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, Huan Ling, et al. 2024. L4GM: Large 4D Gaussian Reconstruction Model. In *Advances in Neural Information Processing Systems*, Vol. 37. 56828–56858.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. 2024. ZeroNVS: Zero-Shot 360-Degree View Synthesis from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9420–9429.

Maximilian Seitzer, Sjoerd van Steenkiste, Thomas Kipf, Klaus Greff, and Mehdi SM Sajjadi. 2023. DyST: Towards Dynamic Neural Scene Representations on Real-World Videos. In *The Twelfth International Conference on Learning Representations*.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. 2023. MVDream: Multi-view Diffusion for 3D Generation. In *The Twelfth International Conference on Learning Representations*.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2023a. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *The Eleventh International Conference on Learning Representations*.

Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. 2023b. Text-to-4D dynamic scene generation. In *Proceedings of the 40th International Conference on Machine Learning*. 31915–31929.

Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. 2023. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818* (2023).

Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. 2024. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10208–10217.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. 2024a. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*. Springer, 1–18.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2024b. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. In *The Twelfth International Conference on Learning Representations*.

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. 2023. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22819–22829.

Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. 2022. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems* 35 (2022), 23371–23385.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. 2024. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*. Springer, 439–457.

Xinzhou Wang, Yikai Wang, Junliang Ye, Fuchun Sun, Zhengyi Wang, Ling Wang, Pengkun Liu, Kai Sun, Xintong Wang, Wende Xie, et al. 2024. Animatabledreamer: Text-guided non-rigid 3d model generation and reconstruction with canonical score distillation. In *European Conference on Computer Vision*. Springer, 321–339.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshops*.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2023. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems* 36 (2023), 8406–8441.

Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, CL Chen, and Lei Zhang. 2023. Consistent123: Improve consistency for one image to 3d object synthesis. *arXiv preprint arXiv:2310.08092* (2023).

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

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. 2024. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*. Springer, 399–417.

Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. 2024. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6796–6807.

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

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. 2024. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048* (2024).

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

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

Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. 2023. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603* (2023).

Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. 2024. A unified approach for text-and image-guided 4d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7300–7309.

Linqi Zhou, Andy Shih, Chenlin Meng, and Stefano Ermon. 2024. Dreampropeller: Supercharge text-to-3d generation with parallel sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4610–4619.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. 2024. Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance. In *European Conference on Computer Vision*. 145–162.

# Splat4D: Diffusion-Enhanced 4D Gaussian Splatting for Temporally and Spatially Consistent Content Creation
## –*Supplementary Material*–

## A Details for 4D Human Generation

The process begins with a source video containing the desired motion for transfer. A pose extraction model [Kocabas et al. 2020] is used to capture SMPL [Loper et al. 2023] sequences, key body landmarks, skeletal poses, and motion trajectories over time. This model processes each frame to extract temporal motion data while preserving dynamic details such as joint rotations, limb movements, and fine-grained motion nuances.

Once the SMPL motion data is obtained, it is projected onto four orthogonal views to serve as input for multi-view generation via the MV-Adapter [Huang et al. 2024]. The projections include sequences of 2D depth images, normal maps, human joint images, and semantic segmentation images. These multi-view motion representations are then fed into the Champ [Zhu et al. 2024] 2D motion transfer model, which generates four sequences of motion images. This model takes static images of the target subject as input and, guided by the extracted pose data, produces motion sequences that replicate the source motion while preserving the visual identity and appearance of the target subject. These sequences serve as an intermediate representation, bridging the source motion and the final 4D output.

The generated 2D motion sequences, denoted as $\{I_t | t \in [1, T]\}$, are subsequently used to initialize a 4D Gaussian field in our framework. Each frame at time $t$ is represented as a set of Gaussians $\mathcal{G}(\mathcal{S}, t) = [\mathcal{X}_t, s_t, r_t, \sigma_t, \zeta_t]$. Using our U-Net-based refinement and Gaussian splatting pipeline, the initial field is optimized to ensure temporal and spatial consistency across views. This process captures intricate details such as loose clothing and rapid movements, resulting in a coherent and realistic 4D human representation.

## B Additional Ablation Study

In Table 5, we perform an ablation study focusing on multiple factors: the inclusion of the feedback loop, the image enhancer, the video diffusion model condition, the use of MV-Adapter compared to SV4D [Xie et al. 2024], as detailed in Section 3.2. The experiments are conducted on Consistent4D [Jiang et al. 2023]. Results demonstrate that using feedback loop optimization, the image enhancer improves visual quality. Condition video diffusion model with first and last frames of input sequence can enhance generation performance. Furthermore, replacing SV4D with MV-Adapter leads to better reconstruction results. These findings highlight the importance of both components in achieving high-quality dynamic scene generation. As discussed in Section 3.2, we illustrate the images before and after the image enhancer process in Fig. 7.

## C Test on Complex Scenarios

In Fig.5 of the main paper, our method successfully handles dancing humans with non-rigid loose clothing. In order to measure the performance on more complex domains, we test on fluids ("splashing water" from Consistent4D [Jiang et al. 2023]) and multi-object interactions ("bouncing ball" from DNeRF [Pumarola et al. 2021],

Table 5. **Additional Ablation Study.** The first row illustrates the results without image enhancer. The second row shows the results using SV4D over MV-Adapter for multi-view generation.

| Model | LPIPS↓ | CLIP-S↑ | FVD-F↓ | FVD-V↓ |
|---|---|---|---|---|
| w/o loop | 0.120 | 0.90 | 831.84 | 473.91 |
| w/o enhancer | 0.108 | 0.96 | 463.39 | 384.26 |
| *SV4D** | 0.105 | 0.94 | 441.72 | 356.25 |
| w/o cond | 0.101 | 0.94 | 425.72 | 339.05 |
| **Ours** | **0.090** | **0.98** | **390.85** | **282.79** |

Table 6. **Evaluation on Liquid Case.**

| Model | CLIP↑ | LPIPS↓ | FVD↓ |
|---|---|---|---|
| Consistent4D [Jiang et al. 2023] | 0.78 | 0.166 | 1282.5 |
| STAG4D [Zeng et al. 2024] | 0.80 | 0.160 | 1231.9 |
| SV4D [Xie et al. 2024] | 0.88 | 0.147 | 772.6 |
| **Ours** | **0.93** | **0.127** | **493.0** |

Table 7. **Evaluation on Multi-object Case.**

| Model | CLIP↑ | LPIPS↓ | FVD↓ |
|---|---|---|---|
| Consistent4D [Jiang et al. 2023] | 0.82 | 0.138 | 1190.7 |
| STAG4D [Zeng et al. 2024] | 0.85 | 0.144 | 1028.2 |
| SV4D [Xie et al. 2024] | 0.90 | 0.125 | 722.3 |
| **Ours** | **0.96** | **0.102** | **428.5** |


Fig. 7. **Effect of Image Enhancer.** We show the difference between human images before and after enhancement with the image enhancer. The enhanced images contain more fine details.

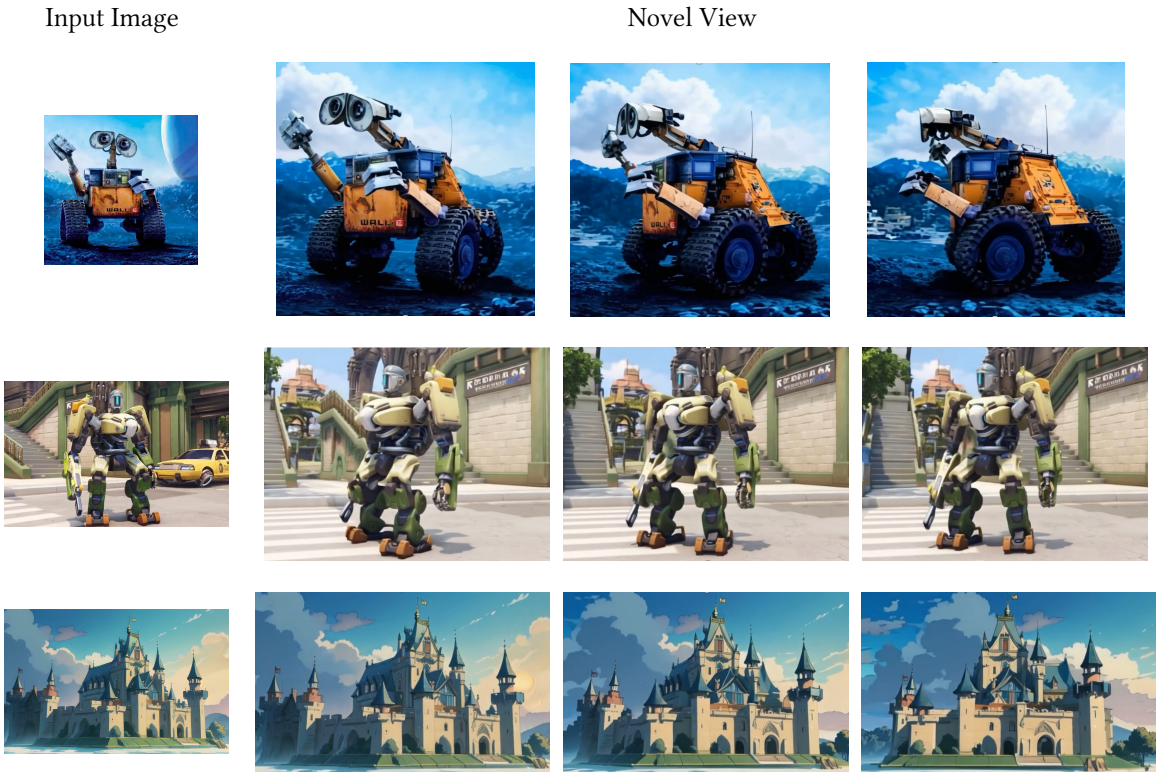


Fig. 8. **4D Generation Results with Background.** The first column on the left shows the input images, while the following three columns show the generation results with a bounded background.

front-view video). The quantitative results for the liquid and multi-object scenario are reported below. Experimental results show that our model can handle complex scenarios such as cloth, fluids and multi-object interactions.

## D More Visualizations

In Fig. 7, we show the difference between original image enhanced image. In Fig. 8 we show our model can handle more challenging cases with bounded 4D scene. In Fig. 9 and Fig. 10 we show more text/image to 4D generation results.

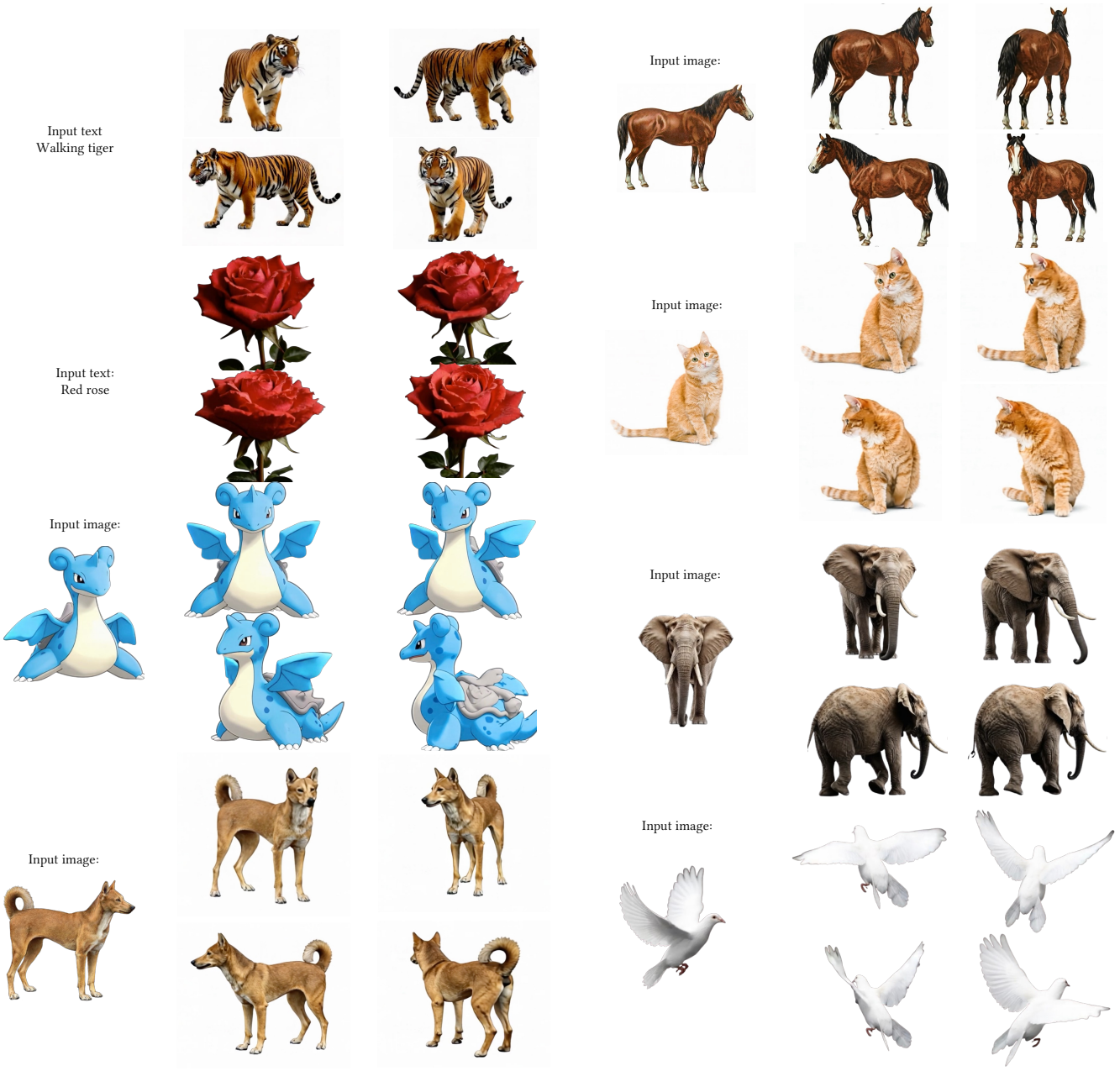


Fig. 9. **4D Generation Results Conditioned on Text/Image.** We present additional text/image-to-4D generation results. On the left are the inputs, and on the right are the generated 4D objects.


Fig. 10. **Video-to-4D Generation**. The left column shows the rendered image of the input view for the 4D object. The right columns show rendered images of novel views.