4D Object-Mover: Distilling Pretrained Diffusion Priors for Object Animation

Tuan Tran Anh[1,2] Jan Eric Lenssen[2] Gerard Pons-Moll[3,2] Julian Chibane[3,2]

[1]German Research Center for Artificial Intelligence
[2] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[3] University of Tübingen, Tübingen AI Center, Germany

https://virtualhumans.mpi-inf.mpg.de/4D_Object_Mover

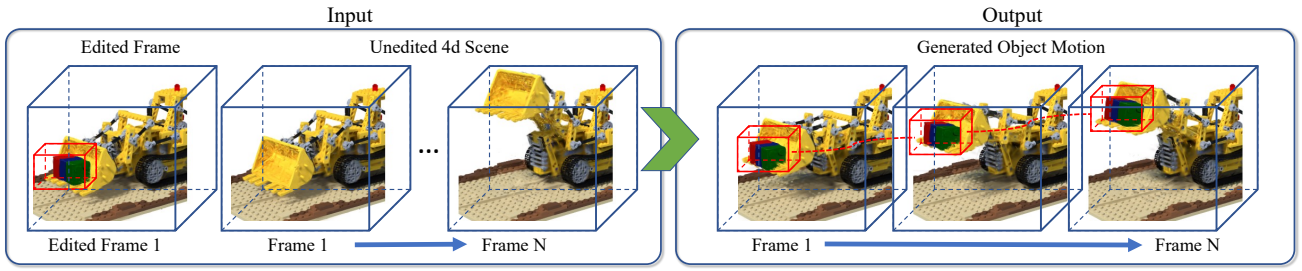

Figure 1. ***Novel Object Motion Generation.*** *Given a 4D scene and an edited first frame with a new object (left), we generate plausible motion for the object in subsequent frames (right).*

Abstract

Recent advances in dynamic scene reconstruction using Neural Radiance Fields (NeRFs) and Gaussian Splatting (GS) have created a demand for effective 4D editing tools. While existing methods primarily focus on appearance alterations or object removal, the challenge of adding objects to 4D scenes, which requires understanding of objects' interactions with the original scene, remains largely unexplored. We present a novel approach to address this gap, focusing on generating plausible motion for newly added objects in 4D scenes. Our key finding is that 2D image-based diffusion models carry strong scene interaction priors that can be extracted from a static scene-object frame and propagated to novel frames of a dynamic 3D scene. Concretely, our method takes an object and its initial placement in a single frame as input, aiming to generate its position and orientation throughout the entire sequence. We first capture the object's appearance, shape, and interaction with the original scene from the static edited frame via fine-tuning a 2D diffusion-based editor. Building on this, we propose an iterative algorithm that leverages the fine-tuned diffusion model to generate frame-to-frame motion for the new object. We show that our method significantly improves 4D motion generation for the new objects compared to prior works on the diverse D-NeRF scene dataset.

1. Introduction

Dynamic view synthesis techniques aim to reconstruct dynamic 3D scenes from captured videos, enabling free-viewpoint and immersive virtual playback. By synthesizing photorealistic novel-view images through rendering, Neural Radiance Fields (NeRF) [35], Gaussian Splatting [26], and their variants have become the leading representations for 3D [3, 4, 19, 69] and 4D dynamic scenes [27, 32, 41, 56]. Beyond mere scene representation, there is a growing interest in creating new, varied scenes derived from original scenes via scene editing. The ability to edit these 3D and 4D representations has significant real-world applications, particularly in the gaming industry, virtual reality, and robotics. Numerous studies have addressed different 3D and 4D scene editing challenges, such as adding objects [10, 16, 30, 36, 45, 61], removing objects, or altering appearances [20].

In 3D, many works allow adding geometry to the original scene [10, 16, 30, 36, 45, 61], but the problem of generating motion for the new object to match the original scene motion is not well-studied. For 4D scenes, previous works have learned to predict an object and its six degree of freedom (6DoF) pose given the input of scanned human point clouds [40]. They demonstrated that they can generate 4D motion of the object when the input is a sequence of point clouds. However, their work is limited to human-object interaction and the type of objects is constrained by the training data. For adding objects to generic 4D scenes, understanding how the added object interacts with the scene to generate motion is crucial. Our work pioneers in this area.

The object-adding problem in 4D scenes can be broken down into two main stages: adding the object to a static frame and generating the motion for the added object. While many works address the former [10, 16, 30, 36, 45, 61], there is a lack of research on the latter. One might attempt to use scene flow and geometry from the original 4D scene to move the object. However, as shown in Fig. 2, even high-quality 4D representations like 4DGS [56] lack reliable scene flow, and tracking methods like Dynamic3DGS [32] may not work out of the box. In this work, we focus on the challenge of generating motion for a new object in a given 4D scene. Our

key finding is that 2D image-based diffusion models carry strong scene interaction priors. We show how to extract them from a static scene-object frame and propagate them to novel frames of a dynamic 3D scene.

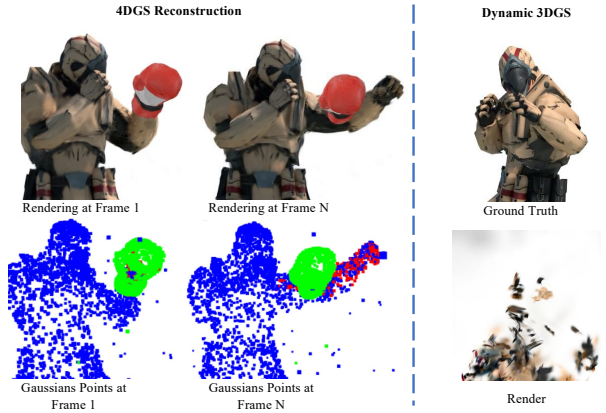

Figure 2. ***Challenge.*** *(Left) Most 4D representations achieve high-quality novel view reconstruction but fail to ensure accurate scene flow or surface geometry, leading to noisy, inconsistent point clouds. We illustrate a naive approach where a soldier's scene flow (blue) moves a novel object (green, a boxing glove) by attaching it to nearby points (red), resulting in incorrect motion and highlighting the need for our scene flow–agnostic approach. (Right) While dense 3D tracking methods exist, they require costly multi-view setups, do not work out of the box, and do not guarantee good novel view reconstruction.*

At a high level, we first aim to *learn the new object's shape, appearance and interaction* from the static edited scene via a pretrained diffusion model. Using the learned information and the 2D priors from the diffusion model, we then *predict the object's six degrees of freedom* (6DoF) information in subsequent frames. An illustration of our approach is shown in Fig. 3.

Two-Stages Motion Generation: 2D generative diffusion models have a strong understanding of complex multi-object interactions [15, 18, 62] and have been adopted for 2D image manipulation [7, 47, 67, 68]. However, while 2D diffusion models can theoretically generate object motion, maintaining consistent appearance and interaction across frames is challenging, as demonstrated in our experiments. To address this, our Stage 1 fine-tunes a 2D diffusion-based editor to capture the object's shape, appearance, and interactions from a single static edited frame (Fig. 3, Stage 1). The fine-tuned editor then generates consistent 2D images of the object interacting with the scene in subsequent frames. In Stage 2 (Fig. 3, Stage 2), we lift these 2D edits to 4D by: 1) initializing the object's pose in each frame based on its pose in the previous frame, and 2) refining the pose through iterative 4D object lifting and regenerating 2D edits, similar to Instruct-NeRF2NeRF[20]. Our method significantly improves 4D motion generation for new objects compared to prior works, as demonstrated on the diverse D-NeRF scene dataset. Our contributions can be summarized as:

- We find that 2D image diffusion models carry strong scene interaction priors useful for object motion generation.
- We show how to leverage 2D diffusion interaction priors from a static scene-object frame and lift their 2D object generations to 4D via an iterative reconstruction.
- Our experiments demonstrate that our scene flow-agnostic generates more faithful motion compared to prior motion generation pipelines.

2. Related Work

2.1. Dynamic Scene Reconstruction and Editing

To expand the success of NeRF into the temporal domain, researchers have pursued the strategy of modeling scenes in 4D domain with time dimension [29, 38, 39, 41, 57]. Some approaches to include time are, DyNeRF [29] using keyframe-based training, VideoNeRF [57] using a spatiotemporal irradiance field from a single video as well as [13, 38, 39, 41, 53] which introduce scene deformations for multi-view and monocular videos by, e.g., a separate MLP. 4D Gaussian Splatting (4DGS) [32, 56], a successor to 3D Gaussian Splatting (GS) [27], extends GS for dynamic scenes by using a deformation field to model Gaussian motions. These methods enable fast reconstruction and real-time novel view synthesis for dynamic scenes.

Extending the success in 3D editing, there is growing interest in developing algorithms for 4D editing. 4D-editor [23] removes objects by propagating input segmentation masks throughout the 4D representation and replacing the segmented area with content from a pre-trained inpainting model. Control4D [46] learns a 4D GAN [11] from the inconsistent outputs of ControlNet [67] to avoid inconsistent supervision signals for 4D portrait editing. AvatarStudio [34] introduces view-and-time-aware Distillation Sampling to distill information from two different diffusion models. Most diffusion-based 4D editing methods focus on human avatars [34, 46] or are limited to appearance changes [22, 34, 37, 46]. Instruct 4D-to-4D [37] and CTRL-D [22] use I-P2P[7] for editing but primarily target appearance modifications. In contrast, our approach enables both object addition and motion generation. Concurrently, CTRL-D also fine-tunes I-P2P for editing but remains focused on appearance changes. There has been no research specifically addressing the incorporation of new objects into generic 4D representations.

2.2. Generation and Reconstruction of Interactions

Several works focus on the task of human-object interaction reconstruction from different kinds of data sources like single image [58, 60, 66], video [12, 49, 59], and multi-view capturing [5, 24, 50], and synthetization of them as well [8, 21, 33, 52, 54, 70]. Other line of works focus on 3D reconstruction of hand-object interac-

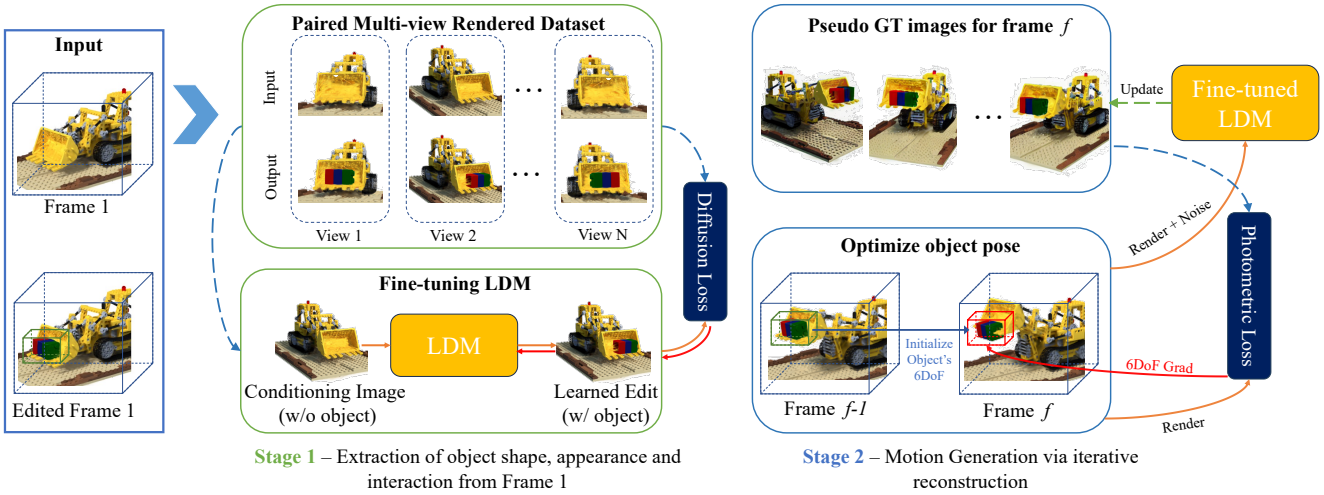

Figure 3. ***Approach:*** *In* ***Stage 1****, we start with the first frame (see* ***Input****,* ***Frame 1****) in our 4D sequence and an edited version including a new object (****Input****,* ***Edited Frame 1****). We create a* ***Paired Multi-view Dataset*** *of rendered images before and after the edit (* ***Stage 1*** *, top). Then, we* ***fine-tune a LDM*** *(latent diffusion model) using this dataset to capture the scene-object interaction. This enables the LDM to accurately place the object into the 2D rendered image of the original scene in any frame. In* ***Stage 2*** *, we generate the object's pose frame by frame. Fist, we* ***inititialze the objects 6DoF*** *pose at frame f by it's pose in frame f-1 (* ***Stage 2*** *, bottom box). Using the fine-tuned editor, we create* ***pseudo GT images for frame f*** *showing the object interacting with the scene for each frame f (* ***Stage 2*** *, top box). Then we optimize the current pose by minimizing the* ***photometric loss*** *(right) between the rendered image and the corresponding pseudo ground truth. Red arrows indicate the backward gradient direction in this process.*

tion [6, 9, 25, 51, 72, 73]. For more general multi-object interaction, Dhamo et al. [14] use an input scene graph where objects are represented as nodes and their relationships as edges to synthesize a set of bounding boxes and object shapes placed at the right arrangement. GraphDreamer [17] generates compositional 3D NeRF scenes from scene graphs by optimizing an SDS loss from a text-to-image diffusion model. Petrov et al. [40] learn from a dataset of human-object interactions to predict an object and its six degrees of freedom (6DoF) pose from scanned human point clouds. However, their work lacks rendering capabilities, is limited to human-object interactions, and is constrained by the objects in their training data. While prior work focuses on human-object interactions or 3D, our approach targets object interactions in general 4D scenes.

2.3. 4D Scene Generation

Video diffusion models have been explored as priors for 4D generation [48]. More recently, image, multiview, and video diffusion models are jointly used as priors for 4D generation [2, 31, 43, 63–65, 71]. Ren et al. and Yin et al. [43, 63] focus on single image to 4D. Most similar to our works are Bahmani et al. and Ling et al. [2, 31] which first synthesize a static 3D scene and then animate the 3D scene with an SDS loss from text-to-video diffusion models. On the other hand, our task has a 4D scene given, and a key challenge is attending to it and understanding the interaction of the novel object with it. Since their model generates images from a text prompt instead of being conditioned on an unedited image as our work, and since video diffusion is hard to fine-tune on a single frame edit, we find their methods fail to produce motion in line with the desired scene.

3. Method

In this section, we introduce the proposed method to generate plausible motion of a new object in a dynamic scene. We begin by introducing the given task in Sec. 3.1, before giving an overview of our method in Sec. 3.2. Then, we provide details of the individual stages of our pipeline in Sec. 3.3 and Sec. 3.4.

3.1. Motion Generation Task

We formulate a novel motion generation task. As input we assume N frames of a 4D scene representation, denoted as $\mathcal{G}_0, \mathcal{G}_1, ..., \mathcal{G}_N$ and a 3D representation $\mathcal{O}$ of an object that is posed within the coordinate space of $\mathcal{G}_0$ by a 6-DOF pose $\mathbf{T}_0 \in \mathbf{SE}(3)$. The goal of our method is to infer poses $\mathbf{T}_1, ..., \mathbf{T}_N \in \mathbf{SE}(3)$ that determine plausible motion of the object over the course of the sequence.

Scene Representations. The given formulation is agnostic to the underlying dynamic scene and object representations, requiring only that they be jointly renderable via a differentiable function R_v, i.e. $\mathbf{I} = R_v(\mathcal{G}, \mathbf{T} \circ \mathcal{O})$, given arbitrary camera pose v, and that $\mathbf{T}$ can be optimized via reconstruction error gradients. Unlike many existing methods, our approach does not rely on accurate scene flow. While we illustrate our method using 3D Gaussian Splatting [27] for the object [27] for the object $\mathcal{O}$ and 4DGS [56] it is gener-

alizable to other representations such as NeRF. Optimization is performed independently for each scene-object pair.

3.2. 4D Object-Mover Overview

To generate motion for a new 3D object within a given 4D scene, two key steps are required: 1) understanding the interaction of the object with the scene, and 2) generating novel motion based on this interaction. Thus, we present a two stage approach. Stage 1 (cf. Sec. 3.3) extracts information about the interaction information between the object and the scene by finetuning a pre-trained latent diffusion model [7] on pairs of original and edited renderings. Stage 2 (cf. Sec. 3.4) uses the extracted information and the pre-trained models priors to sequentially generate the motion for the object frame by frame. Specifically, the fine-tuned diffusion model generates 2D images including novel object poses, which are lifted to 4D by optimizing the object poses over the sequence via a photometric loss.

3.3. Stage 1 - Extracting Object Interaction

The first stage of our pipeline utilizes the given first frame $\mathcal{G}_0$ and the given object $\mathcal{O}$ with pose $\mathbf{T}_0 \in \mathbf{SE}(3)$ to learn the object-scene interaction via fine-tuning. We fine-tune an image-to-image latent diffusion model to map from views of the original scene to edited ones. In the following, we describe our fine-tuning dataset and strategy.

Data Preparation. The dataset consists of paired images captured from the given representation at frame 0, with and without the object. First, we generate a set $\mathcal{V}$ of 360° cameras positioned around the scene with azimuth angles ranging from 0 to 2π. Then, we capture our fine-tuning dataset $\mathcal{D} = \{(R_v(\mathcal{G}_0), R_v(\mathcal{G}_0, \mathbf{T} \circ \mathcal{O}))\}_{v \in \mathcal{V}}$ by rendering original and edited scenes at frame 0 for all $v \in \mathcal{V}$.

Latent Diffusion Model (LDM). Our goal is to use a diffusion model to generate 2D images of an object interacting with the scene in subsequent frames. We achieve this by fine-tuning the model to extract the object's shape, pose, and interaction from the edited static frame. This fine-tuning shifts the model's distribution, enabling accurate synthesis of 2D images with precise placement and realistic motion. We use Instruct-Pix2Pix (I-P2P) [7], a diffusion-based method specialized for image editing. Conditioned on an RGB image c_I and a text-based editing instruction c_T which is processed by text encoder $\mathcal{E}_T(c_T)$, the model takes a noised image (or pure noise) z_t as input and aims to produce z_0 - edited version of c_I based on c_T. In our method, we use the unedited scene's rendered images for c_I, the noised rendered images of the edited scene as z_t, and optimize c_T. Formally, the diffusion model predicts the noise in z_t, using the denoising U-Net ϵ_θ as follows:

$$\hat{\epsilon} = \epsilon_\theta(z_t; t, c_I, \mathcal{E}_T(c_T)) \tag{1}$$

This noise prediction $\hat{\epsilon}$ can be used to derive $\hat{z}_0$, the estimate of the edited image. The denoising process can be queried with a noisy image z_t at any timestep $t \in [0, T]$, with larger t (more noise) produce estimates of $\hat{z}_0$ with more variance, and smaller t values will yield lower variance estimates that adhere more closely to the visible image signal in z_t.

Fine-tuning. For an image $x = R_v(\mathcal{G}_0, \mathbf{T} \circ \mathcal{O})$, the diffusion process adds noise to the encoded latent $z_0 = \mathcal{E}(x)$ producing a noisy latent z_t. We optimize the denoising network ϵ_θ, the text encoder $\mathcal{E}_T()$ and the prompt c_T. Similar to optimizing text-to-image generative diffusion models [44], we minimize the following latent diffusion reconstruction objective:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(x), c_I, \epsilon, t}\Big[\|\epsilon - \epsilon_\theta(z_t, t, c_I, \mathcal{E}_T(c_T)))\|_2^2\Big] \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, 1)$ is the true noise added to the latent representation and $c_I = R_v(\mathcal{G}_0)$ is the conditioning image.

3.4. Stage 2 - Motion Generation

Our goal is to determine the 6DoF pose of the object for each frame. We achieve this by generating 2D pseudo ground truth images of the object interacting with the scene in subsequent frames using the fine-tuned Instruct-Pix2Pix model from Stage 1 (see Fig. 3, Stage 2). The object's motion generation is reformulated as an iterative reconstruction problem, optimizing each frame's pose by minimizing the photometric loss between the rendered image and the pseudo ground truth. This process is iterative, with each frame's pose initialized by the optimized pose from the previous frame. Additionally, we use an iterative dataset update method to regularly update the pseudo ground truth images to stabilize the optimization task.

Iterative Generation: To generate the object's pose for a frame, we first create a set of 2D pseudo ground truth images of the object interacting with the scene in that frame. Formally, for a given frame f, we generate the pseudo ground truth set $D_f = \{I_v\}_{v \in \mathcal{V}}$, where $I_v = U_t(R_v(\mathcal{G}_f, \mathbf{T}_f \circ \mathcal{O}))$ is the output image from the fine-tuned diffusion editor U, obtained by denoising the noised version of the rendered image $R_v(\mathcal{G}_f, \mathbf{T}_f \circ \mathcal{O})$ at noise level t. The rendering of the scene without the object, $R_v(\mathcal{G}_f)$, is used as the conditioning image.

Next we reconstruct the the object pose for frame f using the generated pseudo ground truth D_f. We initialize the current pose with the previous pose $\mathbf{T}_{f-1}$ and find the optimal $\mathbf{T}_f$ by solving the following optimization problem:

$$\begin{aligned}\mathbf{T}_f = \arg\min_{\hat{\mathbf{T}}_f} \sum_{v \in \mathcal{V}} \Big[& SSIM(R_v(\mathcal{G}_f, \hat{\mathbf{T}}_f \circ \mathcal{O}), I_v) \\ & + \|R_v(\mathcal{G}_f, \hat{\mathbf{T}}_f \circ \mathcal{O}) - I_v)\|^2 \Big],\end{aligned} \tag{3}$$

where $SSIM$ represents the structural similarity loss [55].

Iterative Dataset Update. Since the 2D pseudo

ground truth images lack temporal and 3D consistency, optimizing with fully denoised images can cause the object to drift, especially when the target pose is far from the current one due to fast motion. To address this, similar to I-N2N [20], we use images generated by denoising noised versions of the rendered images at varying noise levels, which provide intermediate guidance toward the target pose. We iteratively replace 2D images in the training dataset with those generated by our finetuned LDM, where the noise level determines whether the edited image resembles the current rendering or the input prompt. This process gradually aligns the 3D representation with the LDM's distribution and supports our 6DoF reconstruction task by offering small 2D corrections, making pose adjustment easier. Formally, every $L = 10$ iterations, we replace a pseudo ground truth image $I_v \in D_f$ with $I_v \leftarrow U_t(R_v(\mathcal{G}_f, \mathbf{T}_f \circ \mathcal{O}))$, where t is uniformly sampled from $[t_{min}, t_{max}]$.

Post-processing. Since we use a 2D diffusion model without temporal information, each frame's pseudo ground truth images are not temporally consistent with the previous frames. Therefore, the optimized object motion from the generative motion generation will likely be non-smooth. To address this, apply a smoothing technique to the resulted squence $\mathbf{T}_1, ..., \mathbf{T}_N$ from iterative generation process.

4. Experiments

Dataset Preparation. We use the 4D sequences from the D-NeRF dataset [41], specifically using the 4D reconstructions from [56] For each scene, we sample N frames from the 4D sequence and treat them as independent reconstructions. To edit each 4D scene, we use publicly available 3D assets to obtain the desired object mesh. We create a multiview dataset of the 3D object in Blender, apply Gaussian Splatting for reconstruction, and manually place the object in the first frame of the 4D scene. Implementation details are provided in the supplementary material.

Processing Time. Each scene is fine-tuned with I-P2P for 2400 iterations in Stage 1, followed by 1000 iterations per frame in Stage 2 (32 frames per scene). On average, the entire process takes 4 hours per scene (2.5 hours for Stage 1, 1.5 hours for Stage 2).

Evaluation Metrics. Below we define the metrics we used for evaluating our method:

Directional Clip Similarity (D-CLIP). The Clip model has a strong visual understanding [42]. We measure the mean cosine similarity between two directions of change in the CLIP embedding: from the original to the edited first frame, and from before to after the edit in other frames. Formally, the directional Clip similarity [28] is defined as

$$\text{D-CLIP} = \frac{1}{|V||F|} \sum_{v \in V, f \in F} \text{cos\_sim}(\Delta_f^v, \Delta_0^v), \quad (4)$$

where $\Delta_f^v = \text{Clip}(R_v(\mathcal{G}_f, \mathbf{T}_f \circ \mathcal{O}) - \text{Clip}(R_v(\mathcal{G}_f))$. V and F represent the sets of viewpoints and frames, respectively, and $|\cdot|$ denotes the cardinality of a set.

Chamfer Distance (CD). For each scene we evaluate in this section, we manually prepare a ground truth editing for the sequences' middle and end frame. We measure the mean Chamfer distance (CD) between the predicted points of the object and the ground truth mesh vertices.

4.1. Comparison with Baselines

We compare our method with the following baselines:

Instruct-NeRF2NeRF (I-N2N) [20]. I-N2N is a method for editing static 3D NeRF scenes using text prompts. It iteratively updates the multiview dataset with edited images from the text-conditioned diffusion editor I-P2P, which is then used to reoptimize the NeRF representation. We extend I-N2N to 4D, optimizing the object's poses across all views and frames simultaneously. Unlike the original I-N2N, which modifies scene color and shape, this baseline focuses on the per-frame 6DoF transformation of the object.

Text-to-4D [1, 31]: These works generate 4D scenes from text prompts in two stages: first, a 3D-aware text-to-image diffusion model creates a static 3D scene, then SDS loss from a text-to-video model animates it by guiding a deformation network. Given a static 3D scene with a new object, we adopt this second stage for animation. We explore two variations: T2V-SDS-A, which directly optimizes the per-frame 6DoF object pose using SDS loss, and T2V-SDS-B, which instead models per-point motion with a deformation network, following the original approach.

Gluing. (GLUE) We attach the object to the scene and use the scene's 3D scene flow from 4DGS [56] to move the object. Specifically, for each object point that is in contact with the scene in frame 0 (i.e., within a 3 cm distance from a scene point), we calculate its distance to the nearest scene point. We then adjust the object's pose in later frames to maintain this distance.

Result Analysis. We present our results in Fig. 5 and compare our generated object poses against the baselines and the ground truth in Fig. 4 and Tab. 1. Since we focus on motion, results are best seen in the supplementary video. While baselines with T2V priors can sometimes generate reasonable motions (Fig. 4, T2V-SDS-B, Soldier + Glove), it often fails to generate motion that is consistent with the original scene's motion or distorting the object shape to match the diffusion model's distribution conditioning on the input prompt. This is likely because the original scene's motion and the object fall outside the distribution of the T2V diffusion model. Similarly, I-P2P cannot add the objects to the 2D rendered images in new frames and therefore fails on the motion generation task. While 4DGS [56] achieves impressive 4D scene reconstruction and novel view synthesis results, its resulting scene flows

Method	Bulldozer + Cubes		T-Rex + Wizard hat		Soldier + Glove		Builder + Bucket		Builder + Top hat		Mutant + Axe		Avg.	
	CD↓	D-CLIP↑	CD↓	D-CLIP↑	CD↓	D-CLIP↑	CD↓	D-CLIP↑	CD↓	D-CLIP↑	CD↓	D-CLIP↑	CD↓	D-CLIP↑
T2V-A	70.3	0.36	13.0	0.40	43.8	0.30	67.2	0.26	103.1	0.29	63.4	**0.45**	60.1	0.34
T2V-B	362.8	0.23	14.8	0.22	68.9	0.25	106.2	-0.05	250.1	0.07	61.8	0.44	144.1	0.19
GLUE	14.5	0.41	14.4	0.35	34.0	0.24	5.5	**0.28**	8.8	0.64	9.9	0.44	14.5	0.39
I-N2N	70.1	0.43	16.5	0.44	59.0	0.26	71.2	0.22	93.8	0.21	47.1	0.43	59.61	0.33
Ours	**3.1**	**0.46**	**6.7**	**0.45**	**7.2**	**0.37**	**3.7**	0.27	**8.3**	**0.67**	**9.1**	0.42	**6.35**	**0.44**

Table 1. ***Qualitative Results.*** *We evaluate our method against the baselines and ground truth using Chamfer Distance (CD) in cm and Directional CLIP similarity (D-Clip). Our method significantly outperforms the baseline in both metrics.*

Figure 4. ***Qualitative Comparison****. Illustration showing our results compared to ground truth and other baselines. While we are able to generate consistent interaction with the ground truth, other baselines failed to produce plausible motions.*

can be flawed, as reconstructing 3D scene flow is not its primary objective. As a result, the gluing baseline fails to produce reasonable object motion (eg. Fig 4, Bulldozer + Cubes). Our method, on the other hand, does not require 3D scene flow, can generate plausible motion for the new object in a 4D scene, showing consistent interaction between the original scene and the object, as well as maintaining the interaction from the first edited frame. Similarly, quantitative results significantly outperform baselines in both Chamfer distance and D-Clip metrics (Tab. 1).

4.2. Ablation Study

In this section, we discuss the effectiveness of the design of the two-stage motion generation system:

Without stage 1 (w/o stage 1). In this setting, we use the pretrained default I-P2P model with a manual text instruction describing the scene and interaction, instead of fine-tuning I-P2P in Stage 1.

Without iterative generation (w/o IG). Intead of initializing the objects pose from a fully optimized previous frame, in this setting, we jointly optimize the object's pose for all

Figure 5. ***Qualitative Results.*** *Our method can effectively generates plausible motions for various objects in D-NeRF scenes.*

frames at once. We initialize the pose as given in the first frame. We use the fine-tuned I-P2P from the first stage with the dataset update scheme to generate 2D edited images for pose reconstruction.

Without iterative dataset update (w/o IDU). Our approach generates a pseudo ground truth image dataset using the fine-tuned diffusion model before the optimization of a frame and continuously updates the dataset during the optimization process. In this ablation, we instead only generate the pseudo ground truth once before optimization and start generation from full noise.

We find that without fine-tunning, I-P2P failed to generate images of the scene interacting with the new object (Fig. 7). Hence, the generated motion is implausible. Without the iterative motion generation scheme (w/o IG), the objects fail to follow the motion of the scene (Fig. 6, 3rd row). This occurs because when the desired position of the object in a frame is too far from the current position (e.g. no overlap be-

	full	w/o stage 1	w/o IG	w/o IDU
	Chamfer distance (cm)			
Bulldozer + Cubes	3.1	14.6	64.3	**2.7**
T-Rex + Wizard hat	**6.7**	26.0	16.6	15.8
Soldier + Glove	7.2	17.1	53.5	**6.1**
Builder + Bucket	**3.7**	63.9	63.5	9.0
Builder + Top hat	8.3	27.4	90.8	**6.8**
Mutant + Axe	**9.1**	51.1	51.7	31.1
Mean	**6.3**	33.3	56.7	11.91

Table 2. Ablation results illustrating the importance of each component of our two-stage system. The evaluation metric is the mean Chamfer distance between the object's point cloud and the ground truth mesh's vertices in the middle frame and the last frame. Lower values indicate better results.

Figure 6. ***Ablation Study:*** *Qualitative results showing the impotance of our proposed two-stage pipeline.*

Figure 7. Comparison between outputs from fined-tuned I-P2P (2nd column) and original I-P2P (3rd column). We edit the conditioning image (first row) on a novel frame of each scene. For the original I-P2P we use the manual prompts (from top to bottom order): "A dinosaur fossil wearing a wizard hat", "A humanoid mutant holding an axe", "A lego bulldozer raising red blue green lego blocks". Clearly, I-P2P, without fine-tunning, cannot generate desirable results.

tween target and current pose), photometric losses struggle to provide correct gradients. Without iterative dataset update (w/o IDU), some scenes can generate plausible motion, but the object may drift into non-optimal positions, leading to inaccurate motion in later frames (Fig. 6, 4th row). Generating 2D edits with full noise can produce a scene with the correct object's pose, but the desired pose might be far from the current one. The iterative dataset update scheme generates 2D edits that place the object between the desirable position and the current position, allowing for gradual optimization of the 3D representation with the LDM's distribution. Evaluation against the ground truth suggests that all components are essential for our method (Tab. 2).

5. Conclusion

We present a scene flow–agnostic approach to a novel 4D editing problem: generating motion for a new object in dynamic 4D scenes. Given a 4D scene with an edited first frame containing a newly inserted object, we aim to generate plausible motion for that object. Our key finding is that 2D image-based diffusion models carry strong scene interaction priors that can be extracted from a static scene-object frame and propagated to novel frames of a dynamic 3D scene.

Based on this observation, we proposed a two-stage method: in the first stage, we learn the new object's shape, appearance, and interaction with the original scene by fine-tuning a diffusion-based editor. In the second stage, we iteratively generate the 6DoF motion from one frame to the next by lifting the 2D edited images from the fine-tuned editor to 4d. We demonstrate that our method significantly improves 4D motion generation for new objects compared to prior works on the diverse D-NeRF scene dataset.

We believe our findings can inspire future works in the area of 4D motion generation: distilling motion, and object interaction knowledge into 3D, from models pre-trained on large corpora of 2D images.

Acknowledgements. We thank Verica Lazova and István Sárándi for for their helpful feedback. This work is funded by the Carl Zeiss Foundation. This work is also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans) and the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. G. Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1, Project number 390727645. J. Chibane is a fellow of the Meta Research PhD Fellowship Program - area: AR/VR Human Understanding.

References

[1] 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. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5

[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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024. 3

[3] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *ICCV*, 2021. 1

[4] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 1

[5] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 2

[6] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393. IEEE, 2019. 3

[7] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2, 4

[8] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 387–404. Springer, 2020. 2

[9] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021. 3

[10] Xinhua Cheng, Tianyu Yang, Jianan Wang, Yu Li, Lei Zhang, Jian Zhang, and Li Yuan. Progressive3d: Progressively local editing for text-to-3d content creation with complex semantic prompts. *arXiv preprint arXiv:2310.11784*, 2023. 1

[11] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018. 2

[12] Rishabh Dabral, Soshi Shimada, Arjun Jain, Christian Theobalt, and Vladislav Golyanik. Gravity-aware monocular 3d human-object reconstruction. in 2021 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 12345–12354, 2021. 2

[13] Devikalyan Das, Christopher Wewer, Raza Yunus, Eddy Ilg, and Jan Eric Lenssen. Neural parametric gaussians for monocular non-rigid object reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[14] Helisa Dhamo, Fabian Manhardt, Nassir Navab, and Federico Tombari. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16352–16361, 2021. 3

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[16] Jan-Niklas Dihlmann, Andreas Engelhardt, and Hendrik Lensch. Signerf: Scene integrated generation for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6679–6688, 2024. 1

[17] Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, and Bernhard Schölkopf. Graphdreamer: Compositional 3d scene synthesis from scene graphs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[18] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706, 2022. 2

[19] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *CVPR*, 2024. 1

[20] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19740–19750, 2023. 1, 2, 5

[21] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11374–11384, 2021. 2

[22] Kai He, Chin-Hsuan Wu, and Igor Gilitschenski. Ctrl-d: Controllable dynamic 3d scene editing with personalized 2d diffusion. *arXiv preprint arXiv:2412.01792*, 2024. 2

[23] Dadong Jiang, Zhihui Ke, Xiaobo Zhou, and Xidong Shi. 4d-editor: Interactive object-level editing in dynamic neural radiance fields via 4d semantic segmentation. *arXiv preprint arXiv:2310.16858*, 2023. 2

[24] Yuheng Jiang, Suyi Jiang, Guoxing Sun, Zhuo Su, Kaiwen Guo, Minye Wu, Jingyi Yu, and Lan Xu. Neuralhofusion: Neural volumetric rendering under human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6155–6165, 2022. 2

[25] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 3

[26] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

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

[28] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 5

[29] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022. 2

[30] Yuhan Li, Yishun Dou, Yue Shi, Yu Lei, Xuanhong Chen, Yi Zhang, Peng Zhou, and Bingbing Ni. Focaldreamer: Text-driven 3d editing via focal-fusion assembly. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3279–3287, 2024. 1

[31] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8576–8588, 2024. 3, 5

[32] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024. 1, 2

[33] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 6169–6176. IEEE, 2021. 2

[34] Mohit Mendiratta, Xingang Pan, Mohamed Elgharib, Kartik Teotia, Ayush Tewari, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. Avatarstudio: Text-driven editing of 3d dynamic human head avatars. *ACM Transactions on Graphics (ToG)*, 42(6):1–18, 2023. 2

[35] 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. 1

[36] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Marcus A Brubaker, Jonathan Kelly, Alex Levinshtein, Konstantinos G Derpanis, and Igor Gilitschenski. Reference-guided controllable inpainting of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17815–17825, 2023. 1

[37] Linzhan Mou, Jun-Kun Chen, and Yu-Xiong Wang. Instruct 4d-to-4d: Editing 4d scenes as pseudo-3d scenes using 2d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20176–20185, 2024. 2

[38] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[39] Sungheon Park, Minjung Son, Seokhwan Jang, Young Chun Ahn, Ji-Yeon Kim, and Nahyup Kang. Temporal interpolation is all you need for dynamic neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4212–4221, 2023. 2

[40] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 3

[41] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 5

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[43] 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. 3

[44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 4

[45] Mohamad Shahbazi, Liesbeth Claessens, Michael Niemeyer, Edo Collins, Alessio Tonioni, Luc Van Gool, and Federico Tombari. Inserf: Text-driven generative object insertion in neural 3d scenes. *Arxiv*, 2024. 1

[46] Ruizhi Shao, Jingxiang Sun, Cheng Peng, Zerong Zheng, Boyao Zhou, Hongwen Zhang, and Yebin Liu. Control4d:

Dynamic portrait editing by learning 4d gan from 2d diffusion-based editor. *arXiv preprint arXiv:2305.20082*, 2(6):16, 2023. 2

[47] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024. 2

[48] 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. 3

[49] Zhuo Su, Lan Xu, Dawei Zhong, Zhong Li, Fan Deng, Shuxue Quan, and Lu Fang. Robustfusion: Robust volumetric performance reconstruction under human-object interactions from monocular rgbd stream. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6196–6213, 2022. 2

[50] Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingyi Yu, and Jingya Wang. Neural free-viewpoint performance rendering under complex human-object interactions. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4651–4660, 2021. 2

[51] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 3

[52] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13263–13273, 2022. 2

[53] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2021. 2

[54] Weilin Wan, Lei Yang, Lingjie Liu, Zhuoying Zhang, Ruixing Jia, Yi-King Choi, Jia Pan, Christian Theobalt, Taku Komura, and Wenping Wang. Learn to predict how humans manipulate large-sized objects from interactive motions. *IEEE Robotics and Automation Letters*, 7(2):4702–4709, 2022. 2

[55] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[56] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Wang Xinggang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023. 1, 2, 3, 5

[57] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9421–9431, 2021. 2

[58] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision*, pages 125–145. Springer, 2022. 2

[59] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4757–4768, 2023. 2

[60] Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[61] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *International Conference on Computer Vision (ICCV)*, 2021. 1

[62] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39, 2023. 2

[63] 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. 3

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

[65] Raza Yunus, Jan Eric Lenssen, Michael Niemeyer, Yiyi Liao, Christian Rupprecht, Christian Theobalt, Gerard Pons-Moll, Jia-Bin Huang, Vladislav Golyanik, and Eddy Ilg. Recent Trends in 3D Reconstruction of General Non-Rigid Scenes. *Computer Graphics Forum*, 2024. 3

[66] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 34–51. Springer, 2020. 2

[67] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[68] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9026–9036, 2024. 2

[69] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor:

Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)*, 40(6):1–18, 2021. 1

[70] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision*, pages 518–535. Springer, 2022. 2

[71] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7300–7309, 2024. 3

[72] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 3

[73] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Gears: Local geometry-aware hand-object interaction synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3