AnchorHOI: Zero-shot Generation of 4D Human-Object Interaction via Anchor-based Prior Distillation

Sisi Dai[1], **Kai Xu**[1,2*]

[1]National University of Defense Technology
[2]Institute of AI for Industries (IAII), Chinese Academy of Sciences (CAS)

Abstract

Despite significant progress in text-driven 4D human-object interaction (HOI) generation with supervised methods, the scalability remains limited by the scarcity of large-scale 4D HOI datasets. To overcome this, recent approaches attempt zero-shot 4D HOI generation with pre-trained image diffusion models. However, interaction cues are minimally distilled during the generation process, restricting their applicability across diverse scenarios. In this paper, we propose AnchorHOI, a novel framework that thoroughly exploits hybrid priors by incorporating video diffusion models beyond image diffusion models, advancing 4D HOI generation. Nevertheless, directly optimizing high-dimensional 4D HOI with such priors remains challenging, particularly for human pose and compositional motion. To address this challenge, AnchorHOI introduces an anchor-based prior distillation strategy, which constructs interaction-aware anchors and then leverages them to guide generation in a tractable two-step process. Specifically, two tailored anchors are designed for 4D HOI generation: anchor Neural Radiance Fields (NeRFs) for expressive interaction composition, and anchor keypoints for realistic motion synthesis. Extensive experiments demonstrate that AnchorHOI outperforms previous methods with superior diversity and generalization.

1 Introduction

Humans constantly interact with surrounding objects in daily life, like sitting on a chair, carrying a backpack, or playing a guitar. Text-driven generation of 4D human-object interaction (HOI) is foundational to the 4D virtual world, and has garnered increasing attention for its potential in applications such as AR/VR, video games, embodied AI, and robotics, among many others.

However, generating realistic 4D HOI from natural language remains a challenging task, as it requires extensive prior knowledge to understand both the inherent spatio-temporal complexity and the broad spectrum of interaction types. Existing approaches (Bhatnagar et al. 2022; Diller and Dai 2024; Li et al. 2024) primarily follow the supervised learning paradigm, relying on paired text-HOI data (Li, Wu, and Liu 2023; Bhatnagar et al. 2022; Jiang et al. 2023a) as ground-truth to learn such priors. However, collecting such data at scale is difficult and costly, as it demands sophisticated motion-capture (mocap) for both humans and objects, along with labor-intensive annotations. The limited scale of existing data heavily constrains the scalability and diversity of these supervised approaches.

Recent approaches have taken initial steps toward the zero-shot learning paradigm, aiming to eliminate reliance on paired text-HOI data. While efforts like InterDreamer (Xu et al. 2024b) remove the need for paired text annotations, they still rely on mocap-based HOI data. More recently, methods such as AvatarGO (Cao et al. 2024) attempt to substitute mocap-based data by distilling priors from a pre-trained image diffusion model. However, they focus solely on relative HOI positioning, leaving human deformation and interactive motion with objects unaddressed: (i) the human remains fixed in a canonical pose during interaction composition; (ii) the motion source is derived from a text-to-human motion model without object awareness. While these overlooked aspects are indispensable for realistic 4D HOI generation, they remain unexplored due to the inherent complexity. This calls for both richer priors beyond image diffusion models and more advanced prior distillation techniques for effective guidance.

To this end, we propose AnchorHOI, a novel framework that exploits hybrid priors from pre-trained image and video diffusion models. Given a natural language description as input, AnchorHOI (i) first compose interactions by deeply exploring priors from image diffusion models; (ii) then synthesize motion by leveraging rich motion priors learned from video diffusion models, without the reliance on mocap-based data for either human interaction or motion.

However, achieving expressive interaction composition and realistic motion synthesis is far from straightforward, as two key challenges still stand in the way. ***(i) Adaptive human pose optimization under image diffusion models***. Existing approaches typically fix the human pose during composition, lacking adaptability to interaction-specific scenarios. While adaptive human pose optimization is essential for composing expressive interactions, the complex articulated structure of the human body leads to a high degree of freedom, making such optimization under diffusion priors challenging. ***(ii) Compositional motion extraction from video diffusion models***. While recent video diffusion models demonstrate strong capabilities in generating

*Corresponding author: Kai Xu, kevin.kai.xu@gmail.com

realistic and diverse motion sequences, they often exhibit inter-subject occlusions in compositional scenarios, precisely where grounded human-object contacts occur. This makes it challenging to extract reliable interaction-aware motion for both human and object subjects.

To address these challenges, we introduce a novel anchor-based prior distillation strategy, which circumvents the difficulty of direct optimization under diffusion models. Specifically, through a tractable two-step process: first constructing interaction-aware anchors from textual descriptions, and then leveraging them to guide the target generation. With two tailored anchors, our AnchorHOI incorporates two key innovations as follows: (i) **interaction composition via anchor NeRF**. NeRF, a more effective representation for distilling interaction priors from image diffusion models than complex parametric human models, is thus adopted as our anchor bridge. To alleviate NeRF noise and enhance semantic consistency, we perform pose alignment between the skeletons of the desired human avatar and the anchor NeRF. By composing the posed human avatar with the target object, we achieve thoroughly distilled HOI generation. 2) **motion synthesis via anchor keypoint**. Purely visual cues often miss essential interaction motion information due to occlusion, where keypoints fortunately lie. Consequently, occluded contact keypoints and body keypoints are well suited as anchors. With these anchor keypoints, occluded motion information is reliably recovered, enabling interaction-faithful 4D HOI synthesis.

Our contributions are summarized as follows:

- AnchorHOI takes a further step toward zero-shot text-driven 4D HOI generation, thoroughly exploiting hybrid priors by pioneering anchor-based prior distillation.
- By leveraging anchor NeRFs and anchor keypoints for static interaction composition and dynamic motion synthesis, AnchorHOI circumvents the challenges of high-dimensional optimization and achieves expressive 4D HOI generation.
- Extensive qualitative and quantitative evaluations show that our AnchorHOI substantially outperforms existing methods in both static 3D and dynamic 4D HOI generation.

2 Related Work

3D Content Generation

Benefiting from advances in diffusion-based text-to-image generation (Saharia et al. 2022b,a; Gu, Yang, and Davis 2024; Huang et al. 2019), DreamFusion (Poole et al. 2022) introduced Score Distillation Sampling (SDS) for text-to-3D generation using NeRF, by distilling guidance from pre-trained diffusion models. Subsequent works have improved output quality (Lin et al. 2023; Wang et al. 2023b), controllability (Metzer et al. 2022), and efficiency (Wu et al. 2024), while also exploring textured reconstruction (Richardson et al. 2023; Cao et al. 2023). For 3D humans, methods such as (Kolotouros et al. 2023) generate controllable avatars, though they often require input-specific optimization. Recent approaches like Zero123++ (Shi et al. 2023a) and MV-Dream (Shi et al. 2023b) leverage 2D diffusion models to synthesize consistent multi-view images, serving as inputs for efficient 3D reconstruction (Liu et al. 2023). Large reconstruction models (Hong et al. 2024; Xu et al. 2023) further scale this direction by adopting transformer-based architectures. Despite these advances, generating complex, compositional 3D scenes remains a significant challenge.

3D Compositional Generation

To address the challenge of compositional 3D generation, recent works have explored object layout and relational reasoning. Epstein et al. (Epstein et al. 2024) and GALA3D (Zhou et al. 2024) optimize component arrangements for multi-object scenes. ComboVerse (Chen et al. 2024) introduces spatial-aware SDS to model relations, while GraphDreamer (Gao et al. 2023) leverages large language models to construct object-relation graphs. Despite this progress, modeling human-object interactions remains underexplored. Recently, InterFusion (Dai et al. 2024) generates human-object scenarios by retrieving human poses from offline-constructed, image-reconstructed pose datasets. However, the retrieved poses remain fixed during optimization, limiting adaptability to specific interaction contexts.

4D Content Generation

Recent progress in video diffusion models (Gu et al. 2023) and score distillation sampling has advanced diverse approaches for 4D scene generation. Make-A-Video3D (Singer et al. 2023) adopts HexPlane features for 4D representation. 4D-fy (Bahmani et al. 2023) and DreamGaussian4D (Ren et al. 2023) use multi-stage pipelines to animate static 3D content. Dream-in-4D (Zheng et al. 2023) supports personalized 4D generation via image guidance, while Consistent4D (Jiang et al. 2023b) synthesizes scenes from video input using RIFE (Huang et al. 2022) and super-resolution. 4DGen (Yin et al. 2023) and AnimatableDreamer (Wang et al. 2023a) enable controllable motion via driving videos. More recently, Comp4D (Xu et al. 2024a) and TC4D (Bahmani et al. 2024) introduce trajectory-based generation for compositional 4D scenes. Despite these advances, generating 4D human avatars with realistic object interaction remains challenging. The recent approach AvatarGO (Cao et al. 2024) attempts to address this; however, it lacks human articulation modeling during interaction composition, leading to limited interaction outcomes, such as simple holding.

3 Preliminary Knowledge

SDS. Score Distillation Sampling (SDS), introduced in DreamFusion (Poole et al. 2022), performs iterative optimization to align 3D representations with text-to-image diffusion priors. Compared to non-iterative image generation followed by reconstruction, SDS more reliably distills semantics encoded in image diffusion models, particularly for complex interactions. While $x = g(\Phi)$, x is the 2D image rendered by a differentiable renderer g with model parameters Φ (*e.g.* the MLPs correspondingly in NeRF), under a randomly sampled camera pose. By injecting the sampled

noise ϵ into x at a time step t, the noisy image x_t is produced. The pre-trained 2D text-to-image diffusion model ϕ provides a denoising network $\hat{\epsilon}_\phi(x_t; y, t)$ that predicts the noise $\hat{\epsilon}$ given the noisy image x_t, time step t, and text embedding y. SDS then optimizes the model parameters Φ by minimizing the difference between the predicted noise and the added noise:

$$\nabla_\Phi \mathcal{L}_{\text{SDS}}(\phi, x) = \mathbb{E}_{t,\epsilon}[w(t)(\hat{\epsilon}_\phi(x_t; y, t) - \epsilon)\frac{\partial x}{\partial \Phi}], \quad (1)$$

where $w(t)$ is the weighting term at the time step t.

4 Method

In this section, we begin with problem formulation, followed by anchor illustration and pipeline overview. We then detail our two core components: (i) interaction composition via anchor NeRF, and (ii) motion synthesis via anchor keypoint.

Problem Formulation

AnchorHOI aims to generate dynamic 3D HOI sequences $\mathcal{X}_d$, conditioned on textual input T, *i.e.* $T \rightarrow \mathcal{X}_d$. The input is a natural language description, denoted as $T = \{T_{\text{inter}}, (T_{\text{motion}})\}$, where $T_{\text{inter}} = \{T_{\text{human}}, T_{\text{action}}, T_{\text{object}}\}$ specifies the desired human avatar, interaction type, and object category, respectively. T_{motion} is optional to provide a more detailed motion description. The output is a sequence of 3D HOIs, $\mathcal{X}_d = \{(H_i, O_i)\}_{i=0}^{L-1}$, where H_i and O_i denote the human and object representations at frame i, and L is the total number of frames.

The human representation is defined as $H_i = \{s, r_i^h, t_i^h, M(\theta_i, \Theta)\}$, where s denotes relative scale to the object, r_i^h and t_i^h denote the global rotation and translation, respectively. $M(\theta_i, \Theta)$ represents the human avatar animated by the articulation pose θ_i, which is an explicit mesh defined as $M(\theta_i, \Theta) = \{V, F, C\}$. The posed vertices $V = \mathcal{M}(\theta_i, \psi, \beta, \mathbf{D})$ and faces F are given by the parametric human body model SMPL-X (Pavlakos et al. 2019), and C represents the vertex colors. We denote $\Theta = (\beta, \mathbf{D}, C)$, the parameters for shape sculpting and appearance generation. The object representation is defined as $O_i = \{r_i^o, t_i^o, \Phi\}$, where r_i^o and t_i^o denote the global rotation and translation, Φ represents the object identity including both geometry and appearance.

Following state-of-the-art methods (Cao et al. 2024), we first generate a static 3D HOI instance $\mathcal{X}_s = (H_s, O_s)$, and subsequently extend it to a dynamic 3D HOI sequence $\mathcal{X}_d$, *i.e.*, $T \rightarrow \mathcal{X}_s \rightarrow \mathcal{X}_d$.

Anchor illustration and Pipeline Overview

Anchor illustration. As illustrated in Figure 1, directly distilling priors from image and video diffusion models (IDMs and VDMs) for zero-shot 4D HOI generation, faces two key challenges: *(i)* integrating interaction priors from IDMs into the SMPL-X model (interaction), and *(ii)* transferring 2D interaction motion from VDMs for HOI. Unfortunately, previous work on zero-shot 4D HOI generation avoids these challenges due to the inherent difficulty.

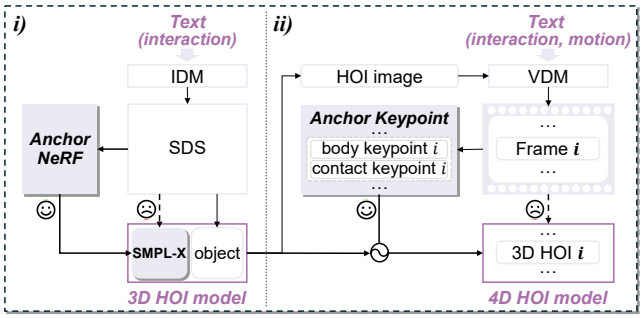

Figure 1: Anchor illustration.

We therefore propose an anchor-based strategy, introducing intermediate anchors to bridge SMPL-X and HOI motion with the priors from IDMs and VDMs. Specifically, two anchors are tailored:

(i) Anchor NeRF. As a visual-prioritized neural representation, NeRF is more adept at distilling rich visual priors from IDMs than the geometry-prioritized SMPL-X model, and thus an ideal anchor bridge. However, visual results from NeRF often suffer from noise and struggle to be directly transferred to the SMPL-X model. We therefore turn to skeleton information rather than pixel-level visual information. Specifically, by aligning the SMPL-X pose with detected skeletons, AnchorHOI enables effective optimization of the SMPL-X structure, thereby capturing reliable interaction cues.

(ii) Anchor keypoints. Visual results of videos generated from VDMs often suffer from inter-subject occlusions in compositional HOI scenarios, missing essential inter-frame differences. In contrast, motion keypoints deliver more robust interaction cues, serving as an ideal anchor bridge. A natural question then arises: *What keypoints are customized anchors for HOI?* To this end, we define anchor keypoints as a combination of body and contact keypoints, offering coherent tracking cues, thus providing a simple yet effective solution for reliable 4D HOI synthesis.

Pipeline overview. As illustrated in Figure 2, the AnchorHOI pipeline comprises two sequential components: (i) interaction composition. We adopt NeRF representations to exploit priors from IDMs and extract human part as anchor NeRF. SMPL-X poses are then optimized to align with skeletons detected from the anchor NeRF, thus capturing interaction cues that are challenging to obtain directly from IDMs. (ii) motion synthesis. We extract both body and contact keypoints from the motion generated by VDMs, treating them as anchor keypoints. The 3D HOI sequences are subsequently optimized to track these anchor keypoints, thereby capturing interaction cues that are otherwise difficult to extract directly from VDMs.

Interaction Composition

This part aims to generate a static 3D HOI instance $\mathcal{X}_s$, *i.e.* $T_{\text{inter}} \rightarrow (H_s, O_s)$ with anchor NeRF based prior distillation from pre-trained IDMs. We first introduce *(i) Anchor NeRF Generation*, and then *(ii) Interaction Generation with*

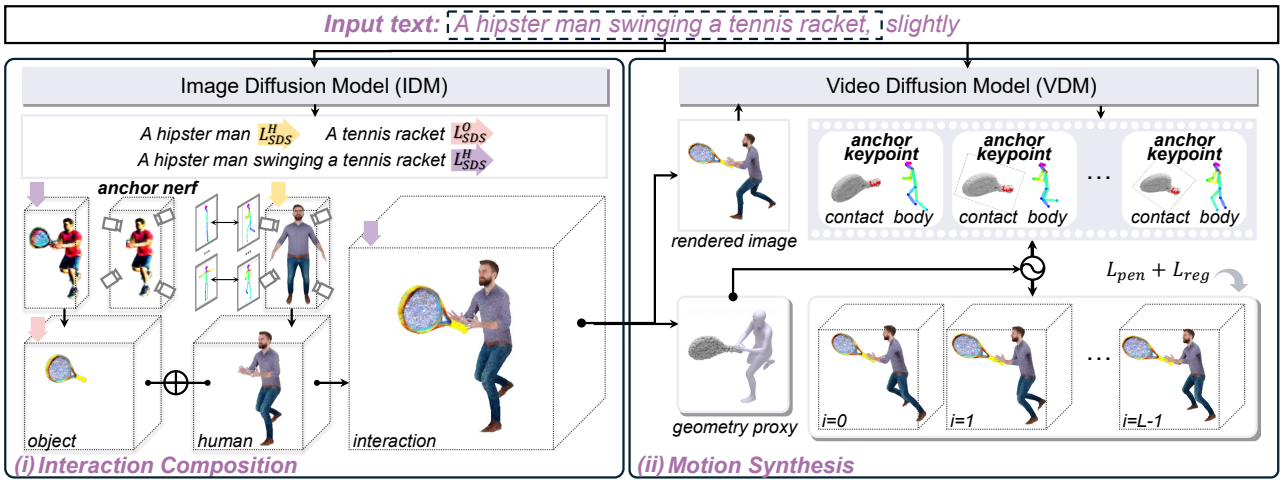

Figure 2: Pipeline overview of AnchorHOI, consisting of two components: (i) interaction composition and (ii) motion synthesis.

Anchor NeRF.

Anchor NeRF generation. We construct an anchor NeRF from a coarse, entangled human-object NeRF optimized under the guidance of IDMs. Specifically, we first obtain a NeRF representation $\hat{\Phi}$ aligned with the textual prompt T_{inter} via SDS optimization: $\nabla_{\hat{\Phi}}\mathcal{L}_{\text{SDS}}(\hat{x})$. We then extract human-isolate NeRF from $\hat{\Phi}$, serving as the anchor NeRF. The extraction is conducted using a multi-view feature alignment loss, formulated as a mean squared error between features: $\nabla_{\hat{\Phi}_A}\mathcal{L}_{\text{align}}(\mathcal{F}(x),\mathcal{F}(\hat{x}))$, where $\mathcal{F}(\cdot)$ denotes the masked RGB feature extractor here.

Interaction generation with anchor NeRF. For human H_s generation, we generate the posed human avatar $H_s = \{s, r_s^h, t_s^h, M(\theta_s, \Theta)\}$ in a two-step manner. To ensure structural completeness and identity consistency with T_{human}, we first generate a canonical avatar using standing pose θ_{stand}, optimizing shape and appearance Θ via SDS optimization: $\nabla_{\Theta}\mathcal{L}_{\text{SDS}}(x)$. This generation process incorporates recent advanced techniques of human avatar generation methods.

We then fit the generated 3D animatable human avatar to the anchor NeRF $\hat{\Phi}_A$, to optimize the remaining parameters $\{s, r_s^h, t_s^h, \theta_s\}$. Specifically, we project canonical views of the anchor NeRF and extract 2D skeleton keypoints using OpenPose (Cao et al. 2019). We then minimize the discrepancy between the projected 3D SMPL-X joints and the detected 2D keypoints across multiple views:

$$\nabla_{\{s,r_s^h,t_s^h,\theta_s\}}\mathcal{L}_{\text{align}} = \sum_{i,j}\rho\big(\Pi(\hat{\mathbf{J}})_j^i - \mathbf{J}_j^i\big), \qquad (2)$$

where $\hat{\mathbf{J}}$ denotes the 3D joint positions of the SMPL-X model, differentiably computed from the model parameters. $\Pi(\cdot)_j^i$ represents the projection of the j-th joint in the i-th camera view, and ρ is the robust Geman-McClure error function (Geman and McClure 1987).

For object generation, the object O_s is first initialized from the segmented object part of Anchor NeRF, and then fully completed via SDS optimization: $\nabla_{\Phi}\mathcal{L}_{\text{SDS}}^{O}(x)$. To preserve the overall human-object interaction, $\nabla_{\Phi}\mathcal{L}_{\text{SDS}}^{I}(x)$ is jointly applied during optimization. Finally, the object mesh is extracted using the marching cubes algorithm (Lorensen and Cline 1998).

Motion Synthesis

Following the interaction composition, this part generates the 4D HOI, *i.e.* $\{T_{\text{inter, (motion)}}, (H_s, O_s)\} \rightarrow \{(H_i, O_i)\}_{i=0}^{L-1}$, with anchor keypoints based prior distillation from pre-trained VDMs. We first introduce *(i) Anchor keypoints extraction*, and then *(ii) Motion optimization with anchor keypoints*.

Anchor keypoints extraction. We extract body and contact keypoints as anchor keypoints from videos generated by the video diffusion model.

(a) Video generation. We adopt a VDM (Zhang and Agrawala 2025) to generate a video $\{F_l^{\text{rgb}}\}_{i=0}^{L-1}$ with L frames, given the textual prompt $T_{\text{inter}}(T_{\text{motion}})$ and the rendered static HOI image I^{rgb}.

(b) Body keypoints. Since inter-subject occlusions often degrade 3D pose estimation, we instead use 2D body keypoints detected by OpenPose (Cao et al. 2019), which offer robust cues even under occlusion. For each frame, OpenPose predicts 18 keypoints (j_i, ω_i), where j_i denotes the normalized pixel coordinates and ω_i the corresponding confidence scores. The keypoints are then reordered to match the SMPL-X joint definitions.

(c) Contact keypoints. While contacts typically occur in occluded regions between subjects, we extract reliable contacts based on the 3D geometric proxy underlying the generated compositional interaction. Specifically, we apply farthest point sampling to obtain a representative subset of surface points from object mesh, denoted as $\mathbf{V}_o \in \mathbb{R}^{N_o\times 3}$. The human vertices used for contact parsing are denoted as $\mathbf{C}(\mathbf{V}_h) \in \mathbb{R}^{N_h\times 3}$, where $\mathbf{C}(\cdot)$ refers to a heuristic selection of SMPL-X mesh vertices as potential contact candidates (Hassan et al. 2019). To ensure the precision of contact

identification, we combine geometric proximity (Bhatnagar et al. 2022; Zhang et al. 2020) and normal alignment (Grady et al. 2021; Yang et al. 2021) constraints, and extract the valid points $\mathcal{C}_{\text{valid}}$ as contact keypoints:

$$\mathcal{C}_{\text{valid}} = \Big\{(i,j) \;\Big|\; (1 - \mathbf{n}_o^i \cdot \mathbf{n}_h^j) < \tau_n, \\ \rho(\|v_o^i - v_h^j\|_2) < \tau_d \Big\}, \tag{3}$$

Here, (i,j) denotes the constructed nearest-neighbor pairs (i,j), where the j-th human vertex $v_j \in \mathbf{C}(\mathbf{V}_h)$ is the closest to the i-th object surface point $v_i \in \mathbf{V}_o$. The normal alignment term $(1 - \mathbf{n}_o^i \cdot \mathbf{n}_h^j)$ encourages the object and the human normals to be antiparallel, thereby favoring physically plausible contacts. This is consistent with the physical fact that interactions occur in contact regions where opposing forces are exerted. The geometric proximity term $\rho(\|v_o^i - v_h^j\|_2) < \tau_d$ encourages physically plausible contacts with close distance. $\rho(\cdot)$ is the Geman-McClure error function (Geman and McClure 1987).

Motion optimization with anchor keypoint. We then transfer the motion from the generated video to 3D HOI by optimizing the human and object representations

$$\left\{\left(\{s, r_i^h, t_i^h, M(\theta_i, \Theta)\}, \{r_i^o, t_i^o, \Phi\}\right)\right\}_{i=0}^{L-1}.$$

Among these, the motion-dependent parameters $\{(r_i^h, t_i^h, \theta_i), (r_i^o, t_i^o)\}$ are optimized, while the motion-invariant variables $\{s, \Theta, \Phi\}$, inherited from the static interaction, are kept fixed and shared across frames.

Taking anchor keypoints as tracking cues, the parameters are then optimized by minimizing the following objective:

$$\mathcal{L}_{\text{total}} = \lambda_{\mathbf{J}}\mathcal{L}_{\mathbf{J}} + \lambda_{\mathbf{C}}\mathcal{L}_{\mathbf{C}} + \lambda_{\text{pen}}\mathcal{L}_{\text{pen}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{4}$$

$$\mathcal{L}_{\mathbf{J}} = \frac{1}{N}\sum_{i=1}^{N} w_i \rho\left(\hat{j}_i - j_i\right) \tag{5}$$

$$\mathcal{L}_{\mathbf{C}} = \frac{1}{|\mathbf{C}|}\sum_{i \in \mathbf{C}} \left\|p_h^i - p_o^i\right\|_2^2 \tag{6}$$

$\mathcal{L}_{\mathbf{J}}$ minimizes the distance between the re-projected SMPL joints and the predicted body keypoints, where w_i is the confidence score of each keypoint, ρ is the Geman-McClure error function (Geman and McClure 1987), and N is the number of keypoints. $\mathcal{L}_{\mathbf{C}}$ minimizes the distance between the i_{th} paired contact keypoint of the human and the object. $\mathcal{L}_{\text{pen}}$ penalizes physical interpenetration between human and object, following (Mihajlovic et al. 2022). The regularization term $\mathcal{L}_{\text{reg}}$ includes the mean squared error between the rendered model frame and video frame, a self-penetration penalty, and a temporal smoothness term.

5 Experiments

To evaluate the effectiveness of AnchorHOI, we conduct a comprehensive comparison with representative methods across 4D visual quality, motion fidelity, and overall interaction plausibility. In addition, we benchmark AnchorHOI against leading static 3D HOI methods to further assess its capability in 3D interaction generation. Experimental results demonstrate that AnchorHOI achieves superior performance in both dynamic 4D and static 3D HOI generation.

Implementation Details

We adopt DeepFloyd (Stability.AI 2023) and a multi-view-consistent image diffusion model (Shi et al. 2023b) to compute SDS gradients, and use the latest video diffusion models (Zhang and Agrawala 2025) to generate 5-second sequences. SMPL-X and VPoser (Pavlakos et al. 2019) serve as human body priors. OpenPose (Cao et al. 2019) is used to extract body joints, and Grounded-SAM (Ren et al. 2024) provides instance segmentation masks. We optimize using Adam (Kingma and Ba 2015) with a learning rate of 0.01, running 3,000 iterations for interaction composition and 1,000 for motion synthesis, on a NVIDIA A6000 GPU.

Experimental Setup

Evaluation baselines. We conduct both quantitative and qualitative evaluations to compare AnchorHOI with representative 4D generation baselines. Specifically, we compare against AvatarGO (Cao et al. 2024), DreamGaussian4D (Ren et al. 2023), and TC4D (Bahmani et al. 2024). Among them, AvatarGO is the most closely related to our approach. However, since the 4D component of AvatarGO has not been publicly released, we include only its static 3D results without motion.

Evaluation metrics. To enable a thorough and reliable quantitative evaluation, we conduct perceptual studies and compute metrics from three complementary perspectives: CLIP score, GPT-4V selection, and perceptual user studies, covering principal quality dimensions such as semantic consistency, physical plausibility, and motion quality.

Qualitative Evaluations

4D HOI comparison. Figure 3, together with Figure 4, presents a comprehensive comparison of 4D HOI generation results. We observe the following: (1) TC4D employs video diffusion models with SDS but treats the entire scene holistically, lacking localized or component-wise motion—an essential element for modeling human-object interactions. (2) DreamGaussian4D, even under powerful guidance (using the same VDM-generated videos as ours), struggles to animate composited HOIs due to its sole reliance on RGB and mask cues. For generating complex HOI dynamics, pixel-level cues alone are insufficient to capture meaningful interactions. (3) AvatarGO cannot be directly compared in 4D due to the unavailable code. However, its static 3D results are fundamentally limited, as it models only position without considering human articulation variance (e.g., humans remain in a fixed canonical pose) while pose variation is necessary for everyday interactions. In the chair example, although AvatarGO can optimize the object position for interaction, the absence of pose modeling prevents it from producing plausible results. Since its 4D outcome directly de-

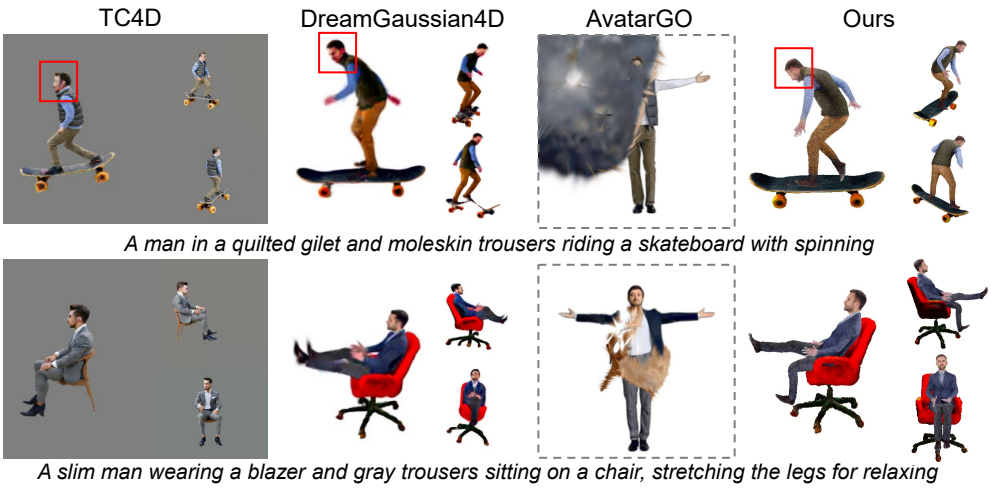

Figure 3: Overall 4D comparison results.

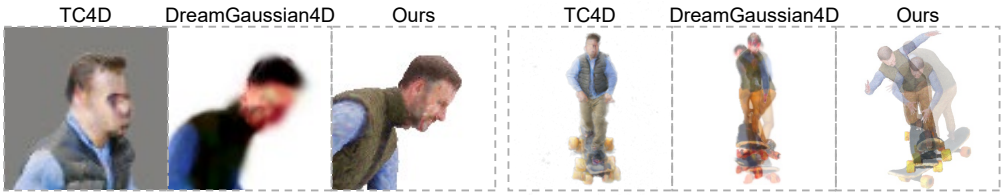

Figure 4: Zoom-in details (left) and motion-centric comparison (right), where opacity conveys motion amplitude: lower opacity indicates larger motion, and vice versa.

	CLIP Score	*GPT-4V Selection (%)*		*User Studies*				
Methods	Semantic ↑	Contact ↑	Overall ↑	Semantic ↑	Contact ↑	Penetration ↑	Motion ↑	Overall ↑
DreamGaussian4D	0.2833	12.50	25.00	2.339	2.382	1.594	2.634	3.339
TC4D	0.3017	16.67	20.83	3.119	2.321	2.863	2.398	3.664
Ours	**0.3149**	**70.83**	**54.17**	**4.794**	**4.756**	**4.673**	**4.874**	**4.833**

Table 1: Quantitative 4D Comparison Results.

pends on the quality of static 3D composition, limitations in 3D modeling inevitably result in implausible 4D outputs.

In contrast, our method achieves more realistic details (e.g., finer visual results in Figure 4(left)), more expressive motion (e.g., larger amplitudes in the skateboard example of Figure 4(right)), and consistent contact awareness throughout the 4D sequences.

3D HOI comparison. Although AnchorHOI targets 4D interaction generation, it also achieves superior performance in static 3D HOI generation. We compare our method with two state-of-the-art baselines, MVDream and InterFusion, as shown in Figure 5. Specifically: (1) MVDream entangles human and object representations, often causing structural artifacts, blurry contacts, or semantically invalid interactions; (2) InterFusion uses fixed, retrieved poses from image reconstructions, lacking adaptability to specific interaction contexts. Benefiting from the proposed anchor NeRF, AnchorHOI produces the most realistic and pose-adaptable

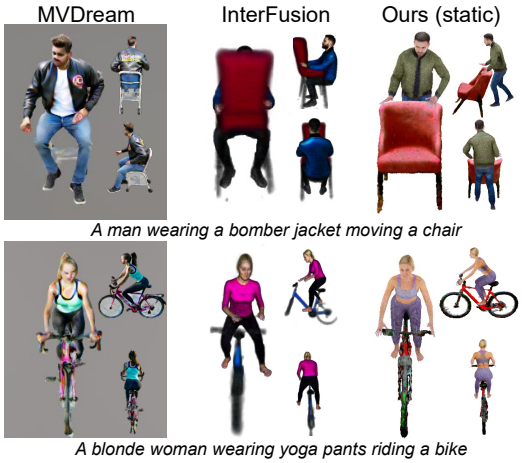

Figure 5: Qualitative 3D comparison results.

(a) *A man with sideburns in a corduroy blazer sitting on a sofa.*

(b) *A Mediterranean man cleaning with a mop.*

Figure 6: Ablation studies on anchor NeRF and anchor keypoints.

Methods	*CLIP Score* Semantic ↑	*GPT-4V Selection (%)* Contact ↑	Overall ↑
MVDream	0.2948	15.79	10.53
InterFusion	0.2951	21.05	26.32
AvatarGO (static)	0.2615	5.26	10.53
Ours (static)	**0.3173**	**57.89**	**52.63**

Table 2: Quantitative 3D comparison results.

Variants	*GPT-4V Selection (%)*
w/o anchor NeRF	5.88
Ours (static)	**94.12**
w/o body keypoints	5.89
w/o contact keypoints	17.65
Ours	**76.47**

Table 3: Quantitative ablation results by GPT-4V.

3D human-object interactions, effectively capturing complex contacts across diverse scenarios. Numeric results in Table 2, including CLIP scores and GPT-4V selections, further demonstrate our superiority.

Quantitative Evaluations

CLIP score. Following common practice, we compute the CLIP score (Radford et al. 2021) to measure the similarity between input text prompts and the corresponding generated results, where a higher score indicates better alignment with input descriptions. AnchorHOI achieves the highest mean CLIP score across evaluation prompts in both 3D and 4D.

GPT-4V selection. Following InterFusion, we further leverage the advanced image understanding capabilities of GPT-4V to enable a more fine-grained evaluation. Specifically, we prompt GPT-4V to (1) select the overall preferred result based on interaction criteria, and (2) assess contact accuracy as a physically grounded metric in isolation. No in-context examples are provided during prompting.

User studies. We further conduct perceptual user studies for 4D evaluation, following (Bylinskii et al. 2023). The numerical results are reported in Table 1.

Ablation Study

Ablation on anchor NeRF. We conduct ablations with and without anchor-NeRF. In the w/o Anchor-NeRF setting, the object and human are generated separately and then composed via SDS directly. As shown in Figure 6 (a), the human pose fails to converge, as position rather than articulation is optimized to satisfy the "sitting on" interaction, resulting in unrealistic results. In contrast, Anchor-NeRF effectively bridges SDS gradients to human articulation, enabling stable convergence and even complex poses (e.g., a crossed-legs posture). This example, along with the quantitative results in Table 3 (up), demonstrates the effectiveness of Anchor-NeRF for faithful interaction composition.

Ablation on anchor keypoints. Figure 6 (b) illustrates the role of anchor keypoints. Body keypoints are essential for capturing human pose; without them, the resulting postures are implausible, even maintaining interaction contacts. Without contact keypoints, the generated results may be visually plausible but lack meaningful physical contact. In contrast, our full setting with both body and contact keypoints produces realistic 4D interaction sequences, as further supported by the quantitative results in Table 3 (down).

6 Conclusion

In this paper, we presented AnchorHOI, a novel framework for zero-shot 4D human-object interaction (HOI) generation with an anchor-based prior distillation strategy. Our approach takes a step forward in interaction-aware 4D generation by effectively leveraging hybrid priors from both image and video diffusion models. Specifically, AnchorHOI tailors anchor NeRFs for interaction composition and anchor keypoints for motion synthesis, enabling effective and reliable prior distillation. Experimental results show that AnchorHOI achieves state-of-the-art performance in both static 3D and dynamic 4D HOI generation.

Limitations and Future Work. One limitation of our method is the current assumption of continuous contact between humans and objects, which overlooks the dynamic nature of contact in real-world scenarios. Another limitation is that our method is currently designed for rigid objects. Extending the framework to accommodate articulated objects with kinematic properties would be a meaningful step toward practical applications.

Acknowledgments

This work was supported in part by the NSFC (62325211, 62132021) and the Major Program of Xiangjiang Laboratory (23XJ01009).

References

Bahmani, S.; Liu, X.; Yifan, W.; Skorokhodov, I.; Rong, V.; Liu, Z.; Liu, X.; Park, J. J.; Tulyakov, S.; Wetzstein, G.; et al. 2024. Tc4d: Trajectory-conditioned text-to-4d generation. In *European Conference on Computer Vision*, 53–72. Springer.

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

Bhatnagar, B. L.; Xie, X.; Petrov, I.; Sminchisescu, C.; Theobalt, C.; and Pons-Moll, G. 2022. BEHAVE: Dataset and Method for Tracking Human Object Interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Bylinskii, Z.; Herman, L.; Hertzmann, A.; Hutka, S.; Zhang, Y.; et al. 2023. Towards better user studies in computer graphics and vision. *Foundations and Trends® in Computer Graphics and Vision*, 15(3): 201–252.

Cao, Y.; Cao, Y.-P.; Han, K.; Shan, Y.; and Wong, K.-Y. K. 2023. DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models. *arXiv preprint arXiv:2304.00916*.

Cao, Y.; Pan, L.; Han, K.; Wong, K.-Y. K.; and Liu, Z. 2024. Avatargo: Zero-shot 4d human-object interaction generation and animation. *arXiv preprint arXiv:2410.07164*.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1): 172–186.

Chen, Y.; Wang, T.; Wu, T.; Pan, X.; Jia, K.; and Liu, Z. 2024. Comboverse: Compositional 3d assets creation using spatially-aware diffusion guidance. *arXiv preprint arXiv:2403.12409*.

Dai, S.; Li, W.; Sun, H.; Huang, H.; Ma, C.; Huang, H.; Xu, K.; and Hu, R. 2024. InterFusion: Text-Driven Generation of 3D Human-Object Interaction. *arXiv preprint arXiv:2403.15612*.

Diller, C.; and Dai, A. 2024. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19888–19901.

Epstein, D.; Poole, B.; Mildenhall, B.; Efros, A. A.; and Holynski, A. 2024. Disentangled 3D Scene Generation with Layout Learning. *arXiv preprint arXiv:2402.16936*.

Gao, G.; Liu, W.; Chen, A.; Geiger, A.; and Schölkopf, B. 2023. GraphDreamer: Compositional 3D Scene Synthesis from Scene Graphs. *arXiv preprint arXiv:2312.00093*.

Geman, S.; and McClure, D. E. 1987. Statistical methods for tomographic image reconstruction. In *Proceedings of the 46th Session of the International Statistical Institute, Bulletin of the ISI*, volume 52.

Grady, P.; Tang, C.; Twigg, C. D.; Vo, M.; Brahmbhatt, S.; and Kemp, C. C. 2021. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1471–1481.

Gu, Z.; Xian, W.; Snavely, N.; and Davis, A. 2023. Factormatte: Redefining video matting for re-composition tasks. *ACM Transactions on Graphics (TOG)*, 42(4): 1–14.

Gu, Z.; Yang, E.; and Davis, A. 2024. Filter-guided diffusion for controllable image generation. In *ACM SIGGRAPH 2024 Conference Papers*, 1–10.

Hassan, M.; Choutas, V.; Tzionas, D.; and Black, M. J. 2019. Resolving 3D human pose ambiguities with 3D scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2282–2292.

Hong, F.; Tang, J.; Cao, Z.; Shi, M.; Wu, T.; Chen, Z.; Wang, T.; Pan, L.; Lin, D.; and Liu, Z. 2024. 3DTopia: Large Text-to-3D Generation Model with Hybrid Diffusion Priors. *arXiv preprint arXiv:2403.02234*.

Huang, Q.; Katsman, I.; He, H.; Gu, Z.; Belongie, S.; and Lim, S.-N. 2019. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4733–4742.

Huang, Z.; Zhang, T.; Heng, W.; Shi, B.; and Zhou, S. 2022. Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision*, 624–642. Springer.

Jiang, N.; Liu, T.; Cao, Z.; Cui, J.; Zhang, Z.; Chen, Y.; Wang, H.; Zhu, Y.; and Huang, S. 2023a. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9365–9376.

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

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Kolotouros, N.; Alldieck, T.; Zanfir, A.; Bazavan, E.; Fieraru, M.; and Sminchisescu, C. 2023. Dreamhuman: Animatable 3d avatars from text. *Advances in neural information processing systems*, 36: 10516–10529.

Li, J.; Clegg, A.; Mottaghi, R.; Wu, J.; Puig, X.; and Liu, C. K. 2024. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, 54–72. Springer.

Li, J.; Wu, J.; and Liu, C. K. 2023. Object Motion Guided Human Motion Synthesis. *ACM Trans. Graph.*, 42(6).

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023.

Magic3D: High-Resolution Text-to-3D Content Creation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2023. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*.

Lorensen, W. E.; and Cline, H. E. 1998. Marching cubes: A high resolution 3D surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, 347–353.

Metzer, G.; Richardson, E.; Patashnik, O.; Giryes, R.; and Cohen-Or, D. 2022. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures. *arXiv preprint arXiv:2211.07600*.

Mihajlovic, M.; Saito, S.; Bansal, A.; Zollhoefer, M.; and Tang, S. 2022. COAP: Compositional articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13201–13210.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D using 2D Diffusion. In *International Conference on Learning Representations*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision.

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

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Richardson, E.; Metzer, G.; Alaluf, Y.; Giryes, R.; and Cohen-Or, D. 2023. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; et al. 2022a. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487*.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022b. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Shi, R.; Chen, H.; Zhang, Z.; Liu, M.; Xu, C.; Wei, X.; Chen, L.; Zeng, C.; and Su, H. 2023a. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*.

Shi, Y.; Wang, P.; Ye, J.; Mai, L.; Li, K.; and Yang, X. 2023b. MVDream: Multi-view Diffusion for 3D Generation. *arXiv:2308.16512*.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2023. Make-a-video: Text-to-video generation without text-video data. In *International Conference on Learning Representations*.

Stability.AI. 2023. Stability AI releases DeepFloyd IF, a powerful text-to-image model that can smartly integrate text into images. https://stability.ai/blog/deepfloyd-if-text-to-image-model.

Wang, X.; Wang, Y.; Ye, J.; Wang, Z.; Sun, F.; Liu, P.; Wang, L.; Sun, K.; Wang, X.; and He, B. 2023a. AnimatableDreamer: Text-Guided Non-rigid 3D Model Generation and Reconstruction with Canonical Score Distillation. *arXiv preprint arXiv:2312.03795*.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023b. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *arXiv preprint arXiv:2305.16213*.

Wu, Z.; Zhou, P.; Yi, X.; Yuan, X.; and Zhang, H. 2024. Consistent3D: Towards Consistent High-Fidelity Text-to-3D Generation with Deterministic Sampling Prior. *arXiv preprint arXiv:2401.09050*.

Xu, D.; Liang, H.; Bhatt, N. P.; Hu, H.; Liang, H.; Plataniotis, K. N.; and Wang, Z. 2024a. Comp4D: LLM-Guided Compositional 4D Scene Generation. *arXiv preprint arXiv:2403.16993*.

Xu, S.; Wang, Y.-X.; Gui, L.; et al. 2024b. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *Advances in Neural Information Processing Systems*, 37: 52858–52890.

Xu, Y.; Tan, H.; Luan, F.; Bi, S.; Wang, P.; Li, J.; Shi, Z.; Sunkavalli, K.; Wetzstein, G.; Xu, Z.; et al. 2023. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217*.

Yang, L.; Zhan, X.; Li, K.; Xu, W.; Li, J.; and Lu, C. 2021. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11097–11106.

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

Zhang, L.; and Agrawala, M. 2025. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*.

Zhang, S.; Zhang, Y.; Ma, Q.; Black, M. J.; and Tang, S. 2020. PLACE: Proximity learning of articulation and contact in 3D environments. In *2020 International Conference on 3D Vision (3DV)*, 642–651. IEEE.

Zheng, Y.; Li, X.; Nagano, K.; Liu, S.; Hilliges, O.; and De Mello, S. 2023. A unified approach for text- and image-guided 4d scene generation. *arXiv preprint arXiv:2311.16854*.

Zhou, X.; Ran, X.; Xiong, Y.; He, J.; Lin, Z.; Wang, Y.; Sun, D.; and Yang, M.-H. 2024. GALA3D: Towards Text-to-3D Complex Scene Generation via Layout-guided Generative Gaussian Splatting. *arXiv preprint arXiv:2402.07207*.